

Two-step estimation of high dimensional additive models

Kengo Kato *

July 24, 2012

Abstract

This paper investigates the two-step estimation of a high dimensional additive regression model, in which the number of nonparametric additive components is potentially larger than the sample size but the number of significant additive components is sufficiently small. The approach investigated consists of two steps. The first step implements the variable selection, typically by the group Lasso, and the second step applies the penalized least squares estimation with Sobolev penalties to the selected additive components. Such a procedure is computationally simple to implement and, in our numerical experiments, works reasonably well. Despite its intuitive nature, the theoretical properties of this two-step procedure have to be carefully analyzed, since the effect of the first step variable selection is random, and generally it may contain redundant additive components and at the same time miss significant additive components. This paper derives a generic performance bound on the two-step estimation procedure allowing for these situations, and studies in detail the overall performance when the first step variable selection is implemented by the group Lasso.

AMS2010 subject classifications: 62G05, 62J99

Key words: additive model, group Lasso, penalized least squares.

1 Introduction

In this paper, we are interested in estimating the nonparametric additive regression model

$$y_i = c^* + g^*(\mathbf{z}_i) + u_i, \quad g^*(\mathbf{z}) = g_1^*(z_1) + \cdots + g_d^*(z_d), \quad \mathbb{E}[u_i | \mathbf{z}_i] = 0, \quad (1.1)$$

*Department of Mathematics, Graduate School of Science, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8526, Japan. Email: kkato@hiroshima-u.ac.jp

where y_i is a dependent variable and $\mathbf{z}_i = (z_{i1}, \dots, z_{id})'$ is a vector of d explanatory variables. Throughout the paper, we assume that the observations are independent and identically distributed (i.i.d.). We presume the situation in which d is larger than the sample size n (in fact we allow for that d is a non-polynomial order in n), but most of g_1^*, \dots, g_d^* are zero functions. So the model of interest is a high dimensional sparse additive model. Let \mathcal{Z} denote the support of \mathbf{z}_1 . Without loosing much generality, we assume that $\mathcal{Z} = [0, 1]^d$. The unknown additive components g_1^*, \dots, g_d^* are known to belong to a given class \mathcal{G} of functions on $[0, 1]$. Throughout the paper, we consider the case in which \mathcal{G} is a Sobolev class: $\mathcal{G} = W_2^\nu([0, 1])$, where ν is a positive integer and

$$W_2^\nu([0, 1]) := \left\{ g : [0, 1] \rightarrow \mathbb{R} : g^{(\nu-1)} \text{ is absolutely continuous such that } \int_0^1 g^{(\nu)}(z)^2 dz < \infty \right\}.$$

For identification of g_1^*, \dots, g_d^* , we assume that

$$\mathbb{E}[g_j^*(z_{1j})] = 0, \quad 1 \leq j \leq d.$$

Let $T^* := \{j \in \{1, \dots, d\} : \mathbb{E}[g_j^*(z_{1j})^2] \neq 0\}$, the index set of nonzero components, and $s^* := |T^*|$, the number of nonzero components. It is assumed that s^* is smaller than n .

There has been a growing interest in estimation of high dimensional sparse additive models (Lin and Zhang, 2006; Ravikumar et al., 2009; Meier et al., 2009; Huang et al., 2010; Koltchinskii and Yuan, 2010; Raskutti et al., 2010; Suzuki et al., 2011; Fan et al., 2011; Buhlmann and van de Geer, 2011). Parallel to parametric regression models, sparsity of the underlying structure makes it possible to estimate consistently the parameter of interest (in this case, the conditional mean function) even when d is larger than n . Estimation accuracy is not a sole goal. In fact, it may happen that, despite the underlying sparsity structure, an estimator containing many redundant components has a good estimation accuracy. However, to make a better interpretation, one wishes to have a concise model. Therefore, the goal is to obtain an estimator that is (i) appropriately sparse, in the sense that it does not contain many redundant additive components, and at the same time (ii) possesses a good estimation accuracy.

A distinctive feature of the present nonparametric estimation, when compared with the parametric case, is that the function class \mathcal{G} is much more complex, which brings a new challenge. To address this problem, in a fundamental paper, Meier et al. (2009), they proposed a penalized least squares estimation method with the novel penalty

term:

$$\tilde{\lambda}_1 \sum_{j=1}^d \sqrt{\|g_j\|_{2,n}^2 + \tilde{\lambda}_2 I(g_j)^2} + \tilde{\lambda}_3 \sum_{j=1}^d I(g_j)^2, \quad (1.2)$$

where

$$\|g_j\|_{2,n}^2 := \frac{1}{n} \sum_{i=1}^n g_j(z_{ij})^2, I(g_j)^2 := \int_0^1 g_j^{(\nu)}(z_j)^2 dz_j.$$

The term $\|g_j\|_{2,n}$ penalizes the event that g_j enters the model, thereby to enforce sparsity of the resulting estimator; the term $I(g_j)$ penalizes roughness of g_j and controls the complexity of the class \mathcal{G} , thereby to avoid an overfitting and guarantee a good estimation accuracy of the resulting estimator. See Koltchinskii and Yuan (2010); Raskutti et al. (2010); Suzuki et al. (2011) for a further progress. An important theoretical fact is that, as Suzuki et al. (2011) showed¹, under suitable regularity conditions such as the uniform boundedness of the error term, the Meier et al. (2009) estimator achieves the minimax rate of convergence (in the L_2 -risk) $s^* \delta^2$ where

$$\delta := \max \left\{ n^{-\nu/(2\nu+1)}, \sqrt{\frac{\log d}{n}} \right\}.$$

See Raskutti et al. (2010) and Suzuki et al. (2011) for minimax rates in our problem. Thus, from a theoretical point of view, their estimator has a good convergence property.

However, we would like to point out that the double penalization strategy that Meier et al. (2009) used may practically lead to a loss of accuracy in estimation/variable selection. In practice, the sparsity penalty brings a shrinkage bias to the selected additive components, so the resulting estimator may have a worse performance than an oracle estimator, which is an “estimator” constructed as if T^* were known, even when the correct model selection is achieved. Furthermore, choosing the tuning parameters in such a way that the estimation accuracy is optimized would result in including too many redundant variables. The problem of shrinkage bias caused by sparsity penalties has been recognized in the parametric regression case. In the linear regression case, Belloni and Chernozhukov (2011b) considered the two-step estimator of the coefficient vector, which corresponds to the least squares estimator applied to the variables selected by the Lasso (Tibshirani, 1996), and observed that in their simulation experiments the two-step estimator significantly outperforms the Lasso estimator because the former can remove a shrinkage bias. Motivated by these observations, we consider the two-step estimation of high dimensional additive models in which the first step implements

¹Suzuki et al. (2011) adopted a slightly different formulation than Meier et al. (2009), i.e., in Suzuki et al. (2011), the Sobolev penalty $I(g_j)$ is replaced by the reproducing kernel Hilbert space norm. However, essentially the same proof applies to the original Meier et al. (2009) estimator.

the variable selection, typically by the group Lasso (Yuan and Lin , 2006), and the second step applies the penalized least squares estimation with Sobolev penalties to the selected additive components. The paper is devoted to a careful study of the theoretical and numerical properties of this two-step estimator.

The main theoretical finding of the paper is to derive a generic bound on the L_2 -risk of the second step estimator. In a typical situation, the bound reduces to

$$\max \left\{ s^* n^{-2\nu/(2\nu+1)}, |\hat{T} \setminus T^*| \delta^2, \left\| \sum_{j \in T^* \setminus \hat{T}} g_j^* \right\|_2^2 \right\}, \quad (1.3)$$

where $\hat{T} \subset \{1, \dots, d\}$ is the index set selected by the first step variable selection. Importantly, this bound applies to any variable selection method such that, roughly speaking, the size $|\hat{T}|$ is stochastically not overly large compared with s^* , and holds in both the situations in which (i) \hat{T} may have redundant variables (i.e., $\hat{T} \setminus T^* \neq \emptyset$), and (ii) \hat{T} may miss significant variables (i.e., $T^* \setminus \hat{T} \neq \emptyset$). This bound has a natural interpretation. The first term $s^* n^{-2\nu/(2\nu+1)}$ corresponds to the oracle rate, the rate that could be achieved when T^* were known; the second term $|\hat{T} \setminus T^*| \delta^2$ corresponds to the effect of selecting redundant variables; the third term $\left\| \sum_{j \in T^* \setminus \hat{T}} g_j^* \right\|_2^2$ corresponds to the effect of missing significant variables.

One may wonder that it is plausible to presume the perfect model selection (i.e., $\hat{T} = T^*$ with probability approaching one), in which case the analysis becomes trivial, since one may guarantee the perfect model selection by applying a hard thresholding method to the first step group Lasso or using the adaptive group Lasso. However, what we need is a bound that applies to a general situation in which \hat{T} may fail to recover T^* . In view of the literature, to guarantee the perfect model selection requires a side condition that the non-zero additive components are well separated from zero (in the L_2 -sense), which is considerably restrictive from a theoretical point of view. In fact, under the side condition, the exact oracle rate $s^* n^{-2\nu/(2\nu+1)}$ will be achievable, and in view of the minimax rate, this means that the complexity of the problem is significantly reduced.² Therefore, to make a meaningful comparison with existing estimators such as the Meier et al. (2009) estimator, one has to establish a performance bound without presuming the perfect model selection. An important aspect of the bound (1.3) is that it characterizes the effect of the first step variable selection in an explicit manner, which makes the analysis non-trivial. Another interesting finding is that, despite the random fluctuation of \hat{T} , the smoothing penalty level in the second step can be taken independent of d .

In this paper, we primarily focus on to use the group Lasso as a first step variable

² $\sqrt{\log d/n}$ can be dominant in δ as long as $\log d/n^{1/(2\nu+1)} \rightarrow \infty$. Our analysis allows for this case.

selection method. The side (and hence not main) contribution of the paper is to establish (some) refined asymptotic results on the statistical properties of the group Lasso estimator for high dimensional additive models, which complements the recent literature on the theoretical study of the group Lasso (Nardi and Rinardo, 2008; Bach, 2008; Wei and Huang, 2010; Huang and Zhang, 2010; Huang et al., 2010; Louinici et al., 2011; Nagahban et al., 2010). The group Lasso, when applied to estimation of additive models, is based on a different idea of dealing with the complexity of function classes, i.e., approximating each function class by a finite dimensional class of functions and controlling the complexity by its dimension.³ Expanding each additive component by a linear combination of given basis functions, selection of additive components reduces to selection of groups of the coefficient vector to the basis expansion, so that the group Lasso turns out to be an effective way of selecting additive components. Combined with the bound (1.3), when the group Lasso is used as a first step procedure, it will be seen that (under suitable regularity conditions, of course) (i) the two-step estimator is at least as good as the Meier et al. (2009) estimator, meaning that it achieves the rate $s^*\delta^2$ in general cases in which \hat{T} may fail to recover T^* (so $\hat{T} \setminus T^* \neq \emptyset$ or $T^* \setminus \hat{T} \neq \emptyset$, or both); (ii) if it happens that the perfect model selection holds, then the two-step estimator enjoys the exact oracle rate $s^*n^{-2\nu/(2\nu+1)}$; (iii) the second step estimation can automatically adapt to both situations (i) and (ii), i.e., adapt to the model selection ability of \hat{T} . We believe that these theoretical results in the context of estimation of high dimensional additive models are useful.

We also carry out simulation experiments to investigate the finite sample property of the two-step estimator. The simulation results suggest that the proposed two-step estimator is a good alternative in estimating high dimensional additive models.

There are a large number of works on the theoretical analysis of penalized estimation methods for high dimensional sparse models, especially on the Lasso for linear regression models (Bunea et al., 2007a,b; Zhao and Yu, 2007; Zhang and Huang, 2008; Meinshausen and Yu, 2009; Wainwright, 2009; Candès and Plan, 2009; Bickel et al., 2009; Zhang, 2009), generalized linear models (van de Geer, 2008; Nagahban et al., 2010) and quantile regression models (Belloni and Chernozhukov, 2011a). See also Buhlmann and van de Geer (2011) for a recent review. In the quantile regression context, Belloni and Chernozhukov (2011a) formally established the theoretical properties of the post-penalized estimator that corresponds to the unpenalized quantile regression estimator applied to the variables selected by the ℓ_1 -penalized estimator. Their

³See Chapter 10 of van de Geer (2000) for two different ideas, namely, penalty and sieve approaches, to deal with the complexity of function classes in nonparametric regression.

analysis is extended to the linear regression case in Belloni and Chernozhukov (2011b). As noted before, the present paper builds on these fundamental papers, but has two important theoretical departures from the previous analysis: (i) the model of interest is a nonparametric additive model, and (ii) the second step estimation has smoothness penalty terms.

The remainder of the paper is organized as follows. Section 2 describes the two-step estimation method. Section 3 presents some simulation experiments. Section 4 is devoted to the theoretical study. Section 5 concludes. Section 6 provides a proof of Theorem 4.1. Some other technical proofs are gathered in Appendices.

Notation: In the theoretical study, we rely on the asymptotic scheme in which d and s^* may diverge as the sample size n . Hence we agree that all parameters values (such as $d, s^* \dots$) are indexed by n and the limit is always taken as $n \rightarrow \infty$, but we omit the index n in most cases. For two sequences $a = a(n)$ and $b = b(n)$, we use the notation $a \lesssim b$ if there exists a positive constant C independent of n such that $a \leq Cb$, $a \asymp b$ if $a \lesssim b$ and $b \lesssim a$, and $a \lesssim_p b$ if $a = O_p(b)$. Let \mathbb{S}^{l-1} denote the unit sphere on \mathbb{R}^l for a positive integer l . Let \mathbf{I}_l denote the $l \times l$ identity matrix. We use $\|\cdot\|_E$ to indicate the Euclidean norm, and let $\|\cdot\|_\infty$ denote the supremum norm. For a matrix \mathbf{A} , let $\|\mathbf{A}\|$ denote the operator norm of \mathbf{A} . For a symmetric positive semidefinite matrix \mathbf{A} , let $\mathbf{A}^{1/2}$ denote the symmetric square root matrix of \mathbf{A} . Let $\|\cdot\|_{2,n}$ and $\|\cdot\|_2$ denote the empirical and population L_2 norms with respect to \mathbf{z}_i 's respectively, i.e., for $g : \mathcal{Z} \rightarrow \mathbb{R}$,

$$\|g\|_{2,n}^2 := \frac{1}{n} \sum_{i=1}^n g(\mathbf{z}_i)^2, \quad \|g\|_2^2 := \mathbb{E}[g(\mathbf{z}_1)^2].$$

To make the notation simpler, if we write the index j in $g_j : [0, 1] \rightarrow \mathbb{R}$, we agree that

$$\|g_j\|_{2,n}^2 := \frac{1}{n} \sum_{i=1}^n g_j(z_{ij})^2, \quad \|g_j\|_2^2 := \mathbb{E}[g_j(z_{1j})^2].$$

2 Two-step estimation

This section describes the proposed estimation method.

First step: Use an appropriate variable selection method to determine a subset \hat{T} of $\{1, \dots, d\}$.

Second step: Apply a penalized least squares method with roughness penalties to the selected additive components:

$$(\tilde{c}, \tilde{g}_j, j \in \hat{T}) := \arg \min_{c \in \mathbb{R}, g_j \in \mathcal{G}, j \in \hat{T}} \left[\frac{1}{2n} \sum_{i=1}^n (y_i - c - \sum_{j \in \hat{T}} g_j(z_{ij}))^2 + \sum_{j \in \hat{T}} \lambda_{2,j}^2 I(g_j)^2 \right], \quad (2.1)$$

subject to the restrictions $\sum_{i=1}^n g_j(z_{ij}) = 0, \forall j \in \hat{T}$, where $\lambda_{2,j} \geq 0$ are smoothing parameter and the term $I(\cdot)$ is the Sobolev penalty:

$$I(f)^2 := \int_0^1 f^{(\nu)}(z)^2 dz.$$

The resulting estimator of g^* is given by $\tilde{g}(z) := \sum_{j \in \hat{T}} \tilde{g}_j(z_j)$. In the theoretical study, to make the argument simple, we let

$$\lambda_{2,1} = \dots = \lambda_{2,d} = \lambda_2.$$

It will be shown that $\lambda_2 \asymp n^{-\nu/(2\nu+1)}$ gives a correct choice.

There are several possible choices for the first step variable selection method. We primarily focus on to use the group Lasso.

Group Lasso: Suppose that we have a set of basis functions $\{\psi_1, \dots, \psi_m\}$ on $[0, 1]$ (except for the constant function). The number $m = m_n$ should be taken such that $m \rightarrow \infty$ as $n \rightarrow \infty$ but $m = o(n)$. It will be shown that $m \asymp n^{1/(2\nu+1)}$ gives an optimal choice. We estimate each additive component by a linear combination of basis functions. Let $\mathcal{G}_m := \{g : [0, 1] \rightarrow \mathbb{R} : g(\cdot) = c + \sum_{k=1}^m \beta_k \psi_k(\cdot), c \in \mathbb{R}, \beta_k \in \mathbb{R}, 1 \leq k \leq m\}$. We consider the estimator:

$$(\hat{c}, \hat{g}_1, \dots, \hat{g}_d) := \arg \min_{c \in \mathbb{R}, g_j \in \mathcal{G}_m, 1 \leq j \leq d} \left[\frac{1}{2n} \sum_{i=1}^n \{y_i - c - \sum_{j=1}^d g_j(z_{ij})\}^2 + \frac{\sqrt{m}\lambda_1}{n} \sum_{j=1}^d \|g_j\|_{2,n} \right], \quad (2.2)$$

subject to the restrictions $\sum_{i=1}^n g_j(z_{ij}) = 0, 1 \leq \forall j \leq d$, where λ_1 is a nonnegative tuning parameter that controls sparsity of the resulting estimator. The resulting estimator of g^* is given by $\hat{g}(z) := \sum_{j=1}^d \hat{g}_j(z_j)$. It will be shown that

$$\lambda_1 \asymp \max \left\{ \sqrt{n}, \sqrt{\frac{n \log d}{m}} \right\}$$

gives a correct choice. Let $\hat{T}^0 := \{j \in \{1, \dots, d\} : \|\hat{g}_j\|_{2,n} > 0\}$.

It is more convenient to concentrate out the constant term when analyzing the estimator \hat{g} . Define $\mathbf{x}_{iG_j} := (\psi_1(z_{ij}), \dots, \psi_m(z_{ij}))'$, $\tilde{\mathbf{x}}_{iG_j} := \mathbf{x}_{iG_j} - \bar{\mathbf{x}}_{G_j}$ ($\bar{\mathbf{x}}_{G_j} := n^{-1} \sum_{i=1}^n \mathbf{x}_{iG_j}$) for $1 \leq j \leq d$ and $\tilde{\mathbf{x}}_i := (\tilde{\mathbf{x}}'_{iG_1}, \tilde{\mathbf{x}}'_{iG_2}, \dots, \tilde{\mathbf{x}}'_{iG_d})'$. Let $\hat{\Sigma}_j := n^{-1} \sum_{i=1}^n \tilde{\mathbf{x}}_{iG_j} \tilde{\mathbf{x}}'_{iG_j}$

for $1 \leq j \leq d$ and $\hat{\Sigma} := n^{-1} \sum_{i=1}^n \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i'$. For $\boldsymbol{\beta} = (\beta_{11}, \dots, \beta_{1m}, \beta_{21}, \dots, \beta_{dm})' \in \mathbb{R}^{dm}$, we use the notation $\boldsymbol{\beta}_{G_j} = (\beta_{j1}, \dots, \beta_{jm})'$ for $1 \leq j \leq d$. Working with this notation, it is seen that $\hat{c} = n^{-1} \sum_{i=1}^n (y_i - \sum_{j=1}^d \hat{g}_j(z_{ij})) = \bar{y} := n^{-1} \sum_{i=1}^n y_i$ and $\hat{g}_j(z_j) = \sum_{k=1}^m \hat{\beta}_{jk}(\psi_k(z_j) - \bar{\psi}_{jk})$ ($\bar{\psi}_{jk} := n^{-1} \sum_{i=1}^n \psi_k(z_{ij})$; $1 \leq j \leq d, 1 \leq k \leq m$), where

$$\hat{\boldsymbol{\beta}} := \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^{dm}} \left[\frac{1}{2n} \sum_{i=1}^n (y_i - \tilde{\mathbf{x}}_i' \boldsymbol{\beta})^2 + \frac{\sqrt{m} \lambda_1}{n} \sum_{j=1}^d \|\hat{\Sigma}_j^{1/2} \boldsymbol{\beta}_{G_j}\|_E \right]. \quad (2.3)$$

Therefore, the estimator \hat{g} is computed by solving the group Lasso problem (2.3), and we call \hat{g} the group Lasso estimator. The group Lasso estimator is known to be groupwise sparse. In the present context, this means that some of additive components are estimated as zero functions.

Some comments are in order.

Remark 2.1. In the group Lasso, there is no need to use common basis functions for all j ; i.e., we may use different basis functions for different j . To make the notation simpler (e.g. to avoid the extra index j to ψ_1, \dots, ψ_m, m and \mathcal{G}_m etc.), we write the group Lasso procedure as it is.

Remark 2.2 (Computation). Because the proposed method is a combination of two commonly used methods, it can be implemented by using standard statistical software packages. In this sense, implementation of the proposed method is simple.

Remark 2.3 (Other options for the first step procedure). Although we primarily focus on the group Lasso for the first step variable selection, it is possible to use other variable selection methods available in the literature. For instance, the nonparametric independence screening (NIS) method proposed in Fan et al. (2011) is known to be a computationally effective way of screening variables. However, in view of (1.3), to obtain a performance bound on the second step estimator, a suitable bound on the magnitude of missed components $\|\sum_{j \in T^* \setminus \hat{T}} g_j^*\|_2^2$ is required, and at the moment it is not known whether NIS ensures a reasonable bound on it. A preferable feature of the group Lasso is that under certain regularity conditions it gives reasonable bounds on both $|\hat{T}^0 \setminus T^*|$ and $\|\sum_{j \in T^* \setminus \hat{T}^0} g_j^*\|_2^2$ (see Section 4).

Remark 2.4 (Other options for the second step procedure). There is an alternative second step estimator of g^* , namely, the sieve least squares estimator applied to the selected additive components:

$$(\check{c}, \check{g}_j, j \in \hat{T}) := \arg \min_{c \in \mathbb{R}, g_j \in \mathcal{G}_m, j \in \hat{T}} \left[\frac{1}{2n} \sum_{i=1}^n \{y_i - c - \sum_{j \in \hat{T}} g_j(z_{ij})\}^2 \right], \quad (2.4)$$

subject to the restrictions $\sum_{i=1}^n g_j(z_{ij}) = 0$, $\forall j \in \hat{T}$ (note: Remark 2.1 applies to this case). Let $\check{g} := \sum_{j \in \hat{T}} \check{g}_j$. It is expected that a similar conclusion (to \tilde{g}) holds for this estimator. In terms of estimation accuracy, it is difficult to judge which is theoretically better. To make the paper focused, we restrict our attention to \tilde{g} and not make a formal study of \check{g} , but compare their finite sample performance by simulations. In our limited simulation experiments, \tilde{g} outperforms \check{g} (see Table 1 ahead), which is a (partial) motivation of studying \tilde{g} .

3 Simulation experiments

This section reports simulation experiments that evaluate the finite sample performance of the estimators. The estimators under consideration are the group Lasso (GL) estimator defined by (2.2), the sieve least square estimator applied to the variables selected by the group Lasso (called GL-SL estimator) defined by (2.4) with $\hat{T} = \hat{T}^0$, the penalized least squares estimator applied to the variables selected by the group Lasso (called GL-PL estimator) defined by (2.1) with $\hat{T} = \hat{T}^0$, the penalized least squares estimator with known true support (called ORACLE estimator) defined by (2.1) with $\hat{T} = T^*$, the Meier et al. (2009) estimator (called MGB estimator). The MGB estimator is defined by a minimizer to the least square criterion function subject to the penalty (1.2) with $\tilde{\lambda}_3 = 0$. The choice $\tilde{\lambda}_3 = 0$ is not theoretically optimal, but what Meier et al. (2009) actually proposed in practice is this estimator, so in these experiments we take $\tilde{\lambda}_3 = 0$.

To implement the group Lasso, we have to determine basis functions and the penalty level λ_1 . We use cubic B-splines with four evenly distributed internal knots (so $m = 7$). To choose the penalty level, we use an AIC type criterion. Let \hat{g}_{λ_1} denote the GL estimator of g^* with penalty level λ_1 . We choose the optimal penalty level that minimizes the criterion

$$\text{AIC}(\lambda_1) = n \log(\sum_{i=1}^n (y_i - \bar{y} - \hat{g}_{\lambda_1})^2 / n) + 2m|\hat{T}_{\lambda_1}^0|,$$

where $\hat{T}_{\lambda_1}^0 := \{j \in \{1, \dots, d\} : \|\hat{g}_{j, \lambda_1}\|_{2, n} \neq 0\}$. Certainly there are other options to choose the penalty level λ_1 , such as the cross validation. Here we use the AIC because of its intuitive nature and since it is simple to implement. To compute GL estimates, we use the package `grplasso` in R. To compute GL-PL estimates and ORACLE estimates, we use the package `mgcv` in R in which the smoothing parameters are automatically optimized according to GCV (by default). See Wood (2006). Comparison with the MGB estimator is not a standard task since its performance depends on the multiple

tuning parameters. To guarantee a fair comparison, according to a preliminary simulation work, we prepared a set of candidate values for $(\tilde{\lambda}_1, \tilde{\lambda}_2)$ and evaluated the performance of the MGB estimator for each $(\tilde{\lambda}_1, \tilde{\lambda}_2)$. The set of candidate values is given by

$$\{(\tilde{\lambda}_1, \tilde{\lambda}_2) : \tilde{\lambda}_1 = \check{\lambda}_1 \times \lambda_{\max}/n, \check{\lambda}_1 \in \{0.12, 0.08, 0.04, 0.02\}, \tilde{\lambda}_2 \in \{0.05, 0.02, 0.01, 0.005\}\},$$

where λ_{\max} is computed by the `lamdamax` option in the `grplasso` package when the minimization problem is transformed to the group Lasso problem.

Each estimator is evaluated by the empirical mean square error (EMSE). Let $\mu_i := c^* + g^*(\mathbf{z}_i)$ and for a generic estimator (\hat{c}, \hat{g}) of (c^*, g^*) , let $\hat{\mu}_i := \hat{c} + \hat{g}(\mathbf{z}_i)$. Then, the EMSE is defined as

$$\text{EMSE} := \mathbb{E}[n^{-1} \sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2].$$

For GL and MGB estimators, we compute the average numbers of numbers of variables selected (NV), false positives (FP) and false negatives (FN).

The number of Monte Carlo repetitions is 500. We consider the case where $n = 400$ and $d = 1,000$. The explanatory variables $\mathbf{z}_i = (z_{i1}, \dots, z_{id})'$ are generated as: $z_{ij} = (w_{ij} + tu_i)/(1 + t)$ for $j = 1, \dots, d$, where $u_i, w_{i1}, \dots, w_{id}$ are i.i.d. uniform random variables on $[0, 1]$. The parameter t controls correlation between variables, i.e., a larger t implies a larger correlation. Three cases $t = 0, 0.5$ or 1 are considered. In what follows, let $g_1(z) = z, g_2(z) = (2z - 1)^2, g_3(z) = \sin(2\pi z)/(2 - \sin(2\pi z))$ and $g_4(z) = 0.1 \sin(2\pi z) + 0.2 \cos(2\pi z) + 0.3 \sin^2(2\pi z) + 0.4 \cos^3(2\pi z) + 0.5 \sin^4(2\pi z)$. We consider two models.

Model 1 $y_i = 5g_1(z_{i1}) + 3g_2(z_{i2}) + 4g_3(z_{i3}) + 6g_4(z_{i4}) + \sqrt{1.74}\epsilon_i, \epsilon_i \sim N(0, 1)$.

Model 2 $y_i = 3.5g_1(z_{i1}) + 2.1g_2(z_{i2}) + 2.8g_3(z_{i3}) + 4.2g_4(z_{i4}) + 3.5g_1(z_{i5}) + 2.1g_2(z_{i6}) + 2.8g_3(z_{i7}) + 4.2g_4(z_{i8}) + \sqrt{1.74}\epsilon_i, \epsilon_i \sim N(0, 1)$.

The coefficients in model 2 are adjusted in such a way that the variance of the conditional mean of y_i given \mathbf{z}_i is roughly the same as in model 1. These designs are essentially adapted from Meier et al. (2009).

The simulation results are given in Tables 1 and 2. Table 2 shows the performance of the MGB estimator with the tuning parameters chosen in such a way that the EMSE is minimized, and hence the EMSEs in Table 2 should be understood as the ideal EMSEs of the MGB estimator. Overall the MGB estimator, with the tuning parameters chosen in such a way that the EMSE is minimized, includes too many

redundant variables. This feature is consistent with the simulation study in Fan et al. (2011).

In model 1, in which the number of nonzero additive components is small ($s^* = 4$) and each nonzero additive component has a relatively large signal, the variable selection by the group Lasso works well, and hence the GL-PL estimator performs strictly better than the ideal MGB estimator in the EMSE in all cases.

In model 2, in which the number of nonzero additive components is large ($s^* = 8$) and each nonzero additive component has a relatively small signal (compared with model 1), the performance of the GL-PL deteriorates, especially when $t = 1$. When $t = 1$, that is, the correlation among \mathbf{z}_i is high, it is difficult to detect the nonzero additive components correctly, and the group Lasso on average does not work very well and the EMSE of the GL-PL estimator is worse than the MGB estimator. However this better performance of the MGB estimator is at the cost of selecting many redundant additive components: on average it includes 75 redundant additive components. It turns out that the performance of the MGB estimator is sensitive to the value of $\tilde{\lambda}_1$ and not to $\tilde{\lambda}_2$. Table 3 shows the performance of the MGB estimator in model 2 with $t = 1$ and $\tilde{\lambda}_2 = 0.05$, and with different values of $\tilde{\lambda}_1$ (the best EMSE among all candidate $(\tilde{\lambda}_1, \tilde{\lambda}_2)$ in model 2 with $t = 1$ is achieved at $\tilde{\lambda}_1 = 0.02 \times \lambda_{\max}/n$, $\tilde{\lambda}_2 = 0.05$, which is the reason why we focus on the $\tilde{\lambda}_2 = 0.05$ case). Increasing $\tilde{\lambda}_1 = 0.02$ to $\tilde{\lambda}_1 = 0.04$ makes the number of false positives small, on average from 75 to 9, but makes the EMSE worse, from 0.528 to 0.921. Taking this into account, we may see that the GL-PL works reasonable well.

4 Theoretical study

4.1 Basic conditions

In this section, we introduce basic conditions commonly used in the analysis of the first and second step estimators.

(C1) (Restriction on the data generating process) $\{(y_i, \mathbf{z}_i')' : i = 1, 2, \dots\}$ are i.i.d. where the pair $(y_1, \mathbf{z}_1')'$ satisfies the model (1.1).

(C2) (Restriction on the (conditional) distribution of u_1) The distribution of u_1 is such that either:

- (a) the support of u_1 is bounded, or;

Table 1: Simulation results

Case	GL				GL-SL	GL-PL	ORACLE
	NV	FP	FN	EMSE	EMSE	EMSE	EMSE
Model 1 ($t = 0$)	5.07	1.07	0.00	1.333	0.398	0.160	0.110
	(1.16)	(1.16)	(0.00)	(0.201)	(0.113)	(0.067)	(0.031)
Model 1 ($t = 0.5$)	5.01	1.02	0.01	0.874	0.236	0.167	0.115
	(1.28)	(1.28)	(0.09)	(0.161)	(0.100)	(0.073)	(0.029)
Model 1 ($t = 1$)	5.25	1.50	0.25	1.083	0.375	0.305	0.120
	(1.96)	(1.72)	(0.52)	(0.234)	(0.239)	(0.247)	(0.031)
Model 2 ($t = 0$)	10.37	2.52	0.15	2.113	0.605	0.332	0.200
	(2.20)	(2.06)	(0.38)	(0.336)	(0.145)	(0.135)	(0.045)
Model 2 ($t = 0.5$)	8.95	1.93	0.97	1.562	0.531	0.425	0.201
	(2.31)	(1.83)	(0.87)	(0.30)	(0.180)	(0.187)	(0.043)
Model 2 ($t = 1$)	6.11	1.24	3.13	1.796	0.996	0.945	0.203
	(1.86)	(1.42)	(0.80)	(0.24)	(0.144)	(0.154)	(0.043)

“GL” refers to the group Lasso estimator, “GL-SL” to the group Lasso + sieve least squares estimator, “GL-PL” to the group Lasso + penalized least squares estimator, “ORACLE” to the penalized least squares estimator with known true support, “NV” to the number of selected variables, “FP” to the false positive, “FN” to the false negative, and “EMSE” refers to the empirical mean square error. Standard deviations are given in parentheses.

Table 2: Simulation results (continued)

Case	MGB			
	NV	FP	FN	EMSE
Model 1 ($t = 0$)	17.86 (6.08)	13.86 (6.08)	0.00 (0.00)	0.357 (0.068)
Model 1 ($t = 0.5$)	13.01 (4.85)	9.01 (4.85)	0.00 (0.00)	0.361 (0.069)
Model 1 ($t = 1$)	70.19 (9.90)	66.19 (9.90)	0.00 (0.00)	0.349 (0.059)
Model 2 ($t = 0$)	62.53 (11.77)	54.53 (11.77)	0.00 (0.00)	0.553 (0.080)
Model 2 ($t = 0.5$)	44.33 (11.11)	36.33 (11.11)	0.00 (0.06)	0.532 (0.081)
Model 2 ($t = 1$)	83.00 (10.14)	75.01 (10.14)	0.01 (0.08)	0.528 (0.070)

“MGB” refers to the (ideal) Meier et al. (2009) estimator, “NV” to the number of selected variables, “FP” to the false positive, “FN” to the false negative, and “EMSE” refers to the empirical mean square error. Standard deviations are given in parentheses.

Table 3: Simulation results for the MGB estimator in model 2 with $t = 1$ and $\tilde{\lambda}_2 = 0.05$, and different values of $\tilde{\lambda}_1$

	$\tilde{\lambda}_1 = 0.12$	$\tilde{\lambda}_1 = 0.08$	$\tilde{\lambda}_1 = 0.04$	$\tilde{\lambda}_1 = 0.02$
NV	5.14 (0.99)	7.82 (1.94)	16.77 (4.51)	83.00 (10.14)
FP	0.27 (0.55)	1.41 (1.59)	8.89 (4.48)	75.01 (10.14)
FN	3.14 (0.78)	1.60 (1.00)	0.12 (0.35)	0.01 (0.08)
EMSE	2.167 (0.146)	1.624 (0.145)	0.921 (0.125)	0.528 (0.070)

“NV” refers to the number of selected variables, “FP” to the false positive, “FN” to the false negative, and “EMSE” refers to the empirical mean square error. Standard deviations are given in parentheses.

- (b) $u_1|z_1 \sim N(0, \sigma_u(z_1)^2)$ and $\sigma_u(z_1) \leq \sigma_u$ almost surely for some constant σ_u independent of n .

(C3) (Restrictions on the distribution of z_1)

- (i) The support of z_1 is $[0, 1]^d$.
- (ii) Let q_j denote the density of z_{1j} for each $1 \leq j \leq d$. Then, q_j is bounded away from zero on $[0, 1]$ uniformly over $1 \leq j \leq d$, i.e., there exists a positive constant c_q such that $c_q \leq q_j$ on $[0, 1]$ for all $1 \leq j \leq d$.

(C4) (Restriction on smoothness of the additive components) $g_j^* \in \mathcal{G}$ for all $j \in T^*$, where $\mathcal{G} = W_2^\nu([0, 1])$ for some positive integer ν .

(C5) (Preliminary restrictions on d and s^*) $d \geq n$, $\log d/n^{2\nu/(2\nu+1)} \rightarrow 0$ and $1 \leq s^* \leq n$.

Condition (C1) is a standard assumption. Condition (C2) needs an explanation. It turns out that the key property to our rate results in Theorems 4.1 and 4.2 (and indeed to those in Koltchinskii and Yuan (2010), Raskutti et al. (2010) and Suzuki et al. (2011) as well) is the normal concentration property (around its mean and given z_1, \dots, z_n) of a random variable of the form $\sup_{t \in \mathcal{T}} \sum_{i=1}^n u_i t_i$ where \mathcal{T} is a bounded and countable subset of \mathbb{R}^n (\mathcal{T} typically depends on z_1, \dots, z_n). In fact, condition (C2) is a primitive sufficient condition that ensures this normal concentration property. See Appendix C for more discussion on this condition. Condition (C3) is standard in the series estimation literature (see e.g. Newey, 1997). Condition (C4) restricts the smoothness property of each additive component g_j^* . We exclude the case that ν is fractional. Condition (C5) is a preliminary restriction on the growth rate of d . To make the technical argument simpler, we here assume that $d \geq n$. Because our primal concern is on the “ $d \gg n$ ” case, this restriction does not bind. The second part of condition (C5) is to restrict d not to grow too fast. The last part of condition (C5) is a natural restriction on s^* .

4.2 A generic bound on the second step estimator

In this section, we present a generic bound on the second step estimator. Although we primarily focus on to use the group Lasso as a first step procedure, the result of this section holds for any variable selection method satisfying the high level condition stated below.

We first prepare some notation. Let $\mathcal{G}_j := \{g_j \in \mathcal{G} : \mathbb{E}[g_j(z_{1j})] = 0\}$. For a subset $T \subset \{1, \dots, d\}$, define

$$\alpha(T) := \inf \left\{ \alpha > 0 : \sum_{j \in T} \|g_j\|_2^2 \leq \alpha \left\| \sum_{j \in T} g_j \right\|_2^2, \forall g_j \in \mathcal{G}_j (j \in T) \right\}.$$

The quantity $\alpha(T)^{-1}$ is an analogue of sparse minimum eigenvalues to the infinite dictionary case. It is clear that when $z_{1j}, j \in T$ are independent, $\alpha(T) = 1$, so $\alpha(T)$ measures the dependence among variables $z_{1j}, j \in T$ (recall that each function in \mathcal{G}_j is centered such that $\mathbb{E}[g_j(z_{1j})] = 0$). Such a quantity appears in other papers on estimation of high dimensional additive models (Koltchinskii and Yuan, 2010; Suzuki et al., 2011). Observe that $\alpha(T) \geq 1$ for any non-empty $T \subset \{1, \dots, d\}$.

We introduce a high level condition on \hat{T} . Put

$$\delta := \delta_n := \max \left\{ n^{-\nu/(2\nu+1)}, \sqrt{\frac{\log d}{n}} \right\}.$$

(C6) (Restriction on the set \hat{T}) $n^{1/2(2\nu+1)} \delta \alpha(T^* \cup \hat{T}) |T^* \cup \hat{T}| = o_p(1)$.

Note that under condition (C5), $n^{1/2(2\nu+1)} \delta = \max\{n^{-(2\nu-1)/2(2\nu+1)}, \sqrt{\log d / n^{2\nu/(2\nu+1)}}\} \rightarrow 0$. Condition (C6) requires that $\alpha(T^* \cup \hat{T})$ and $|T^* \cup \hat{T}|$ are not too large. In the canonical case in which $\alpha(T^* \cup \hat{T}) \lesssim_p 1$ and $|\hat{T}| \lesssim_p s^*$, condition (C6) is satisfied if $s^* = o[\min\{n^{(2\nu-1)/2(2\nu+1)}, \sqrt{n^{2\nu/(2\nu+1)} / \log d}\}]$. We shall comment that, even when T^* were known, a condition analogous to (C6) is needed to obtain a reasonable bound on the estimator, so we believe that, as long as $|\hat{T}|$ is stochastically not overly large compared with s^* , condition (C6) is a reasonable restriction. It will be shown that, when the group Lasso is used as a first step procedure, $|\hat{T}^0| \lesssim_p s^*$.

We are now in position to state the main theorem of this paper.

Theorem 4.1. *Assume conditions (C1)-(C6). Take λ_2 such that $\lambda_2 \geq A_{2,u,\nu} n^{-\nu/(2\nu+1)}$, where $A_{2,u,\nu}$ is some positive constant depending only on the distribution of u_1 and the smoothness index ν . Then, we have*

$$\begin{aligned} & \|\tilde{g} - g^*\|_2^2 + \lambda_2^2 \sum_{j \in \hat{T}} I(\tilde{g}_j)^2 \\ & \lesssim_p \max \left\{ \alpha(T^* \cup \hat{T}) |T^* \cap \hat{T}| n^{-2\nu/(2\nu+1)}, \alpha(T^* \cup \hat{T}) |\hat{T} \setminus T^*| \delta^2, n^{-2\nu/(2\nu+1)} \|g^*\|_2^2 \right. \\ & \quad \left. \lambda_2^2 \sum_{j \in T^* \cap \hat{T}} I(g_j^*)^2, \|\sum_{j \in T^* \setminus \hat{T}} g_j^*\|_2^2, n^{-2\nu/(2\nu+1)} \sum_{j \in T^* \setminus \hat{T}} I(g_j^*)^2 \right\}. \end{aligned}$$

In particular, in the canonical case in which (i) $\alpha(T^* \cup \hat{T}) \lesssim_p 1$; (ii) $\|g^*\|_2^2 \lesssim s^*$; (iii) $\sum_{j \in T^*} I(g_j^*)^2 \lesssim s^*$, for $\lambda_2 \geq A_{2,u,\nu} n^{-\nu/(2\nu+1)}$, we have

$$\|\tilde{g} - g^*\|_2^2 + \lambda_2^2 \sum_{j \in \hat{T}} I(\tilde{g}_j)^2 \lesssim_p \max \left\{ s^* \lambda_2^2, |\hat{T} \setminus T^*| \delta^2, \|\sum_{j \in T^* \setminus \hat{T}} g_j^*\|_2^2 \right\}. \quad (4.1)$$

Remark 4.1. In principle, it is possible to state the theorem in a non-asymptotic manner; however, to make the exposition clear, we state the theorem as it is.

Interestingly, λ_2 can be taken independent of d despite the random fluctuation of \hat{T} . This is in contrast to the fact that, e.g. in Koltchinskii and Yuan (2010); Raskutti et al. (2010), penalty levels (on smoothness) should scale as $\log d$ as $d \rightarrow \infty$.

This theorem characterizes the effect of the first step variable selection in an explicit manner: in (4.1), (i) the first term $s^* \lambda_2^2$ reflects the oracle rate; (ii) the second term $|\hat{T} \setminus T^*| \delta^2$ reflects the cost of selecting redundant components; (iii) the third term $\|\sum_{j \in T^* \setminus \hat{T}} g_j^*\|_2^2$ reflects the magnitude of missed components. We will investigate the behaviors of these terms when the group Lasso is used as a first step procedure.

4.3 Properties of the group Lasso

In this section, we collect the statistical properties (namely the convergence rate and the model selection property) of the group Lasso estimator \hat{g} used as a first step estimator. Although such properties have been well studied in the literature especially for the parametric regression case (Nardi and Rinardo, 2008; Bach, 2008; Ravikumar et al., 2009; Huang and Zhang, 2010; Wei and Huang, 2010; Huang et al., 2010; Louinici et al., 2011; Nagahban et al., 2010), we could not find results that we exactly need in the very present setting, in particular an explicit scaling condition on the triple (d, s^*, m) that guarantees the statistical properties. For the sake of completeness, we state here these properties. Their proofs are found in Appendix.

We begin with introducing restrictions on basis functions.

(C7) (Restrictions on basis functions used in the first step estimation)

- (a) $\sup_{z \in [0,1]} \|(\psi_1(z), \dots, \psi_m(z))'\|_E = O(m^{1/2})$.
- (b) $E[\tilde{\mathbf{x}}_{1G_j} \tilde{\mathbf{x}}'_{1G_j}] = \mathbf{I}_m$ for all $1 \leq j \leq d$.
- (c) $\inf_{g \in \mathcal{G}_m^{T^*}} \|g^* - g\|_2^2 \lesssim s^* m^{-2\nu}$, where $\mathcal{G}_m^{T^*} := \{g : \mathcal{Z} \rightarrow \mathbb{R} : g(\mathbf{z}) = \sum_{j \in T^*} g_j(z_j) (\mathbf{z} = (z_1, \dots, z_d)') , g_j \in \mathcal{G}_m (j \in T^*)\}$.

We refer to Newey (1997) for some basic materials on series estimation. Condition (C7)-(a) is satisfied for splines and Fourier bases. Condition (C7)-(b) is a normalization

condition, and does not lose any generality as long as we are concerned with the analysis of the statistical properties of the group Lasso estimator. Condition (C7)-(c) corresponds to condition (C4) and is thought to be a reasonable restriction. Consider, for instance, ψ_1, \dots, ψ_m are spline functions of degree $(\nu + 1)$ on $[0, 1]$ with equidistant knots. By Corollary 6.26 of Schumaker (2007), there exists a $g^m = \sum_{j \in T^*} g_j^m \in \mathcal{G}_m^{T^*}$ such that $\sum_{j \in T^*} \|g_j^* - g_j^m\|_2^2 \lesssim m^{-2\nu} \sum_{j \in T^*} I(g_j^*)^2$. Because $E[g_j^*(z_{1j})] = 0$, g_j^m may be taken such that $E[g_j^m(z_{1j})] = 0$. Therefore, letting for a subset $T \subset \{1, \dots, d\}$,

$$\beta(T) := \inf \left\{ \beta > 0 : \left\| \sum_{j \in T} g_j \right\|_2^2 \leq \beta \sum_{j \in T} \|g_j\|_2^2, \forall g_j \in \mathcal{G}_j (j \in T) \right\},$$

if $\sum_{j \in T^*} I(g_j)^2 \lesssim s^*$ and $\beta(T^*) \lesssim 1$, then $\|g^* - g^m\|_2^2 \leq \beta(T^*) \sum_{j \in T^*} \|g_j^* - g_j^m\|_2^2 \lesssim \beta(T) m^{-2\nu} \sum_{j \in T^*} I(g_j^*)^2 \lesssim s^* m^{-2\nu}$. The restriction that $\sum_{j \in T^*} I(g_j^*)^2 \lesssim s^*$ is reasonable. Trivial examples in which $\beta(T^*) \lesssim 1$ are the case that $s^* \lesssim 1$ or the case that $z_{1j}, j \in T^*$ are independent. Conditions similar to $\beta(T^*) \lesssim 1$ appear in other papers such as Koltchinskii and Yuan (2010).

We now start to investigate the statistical properties of the group Lasso estimator. To this end, we prepare some notation. Define the event

$$\Omega_0 := \{\|\hat{\Sigma}_j^{1/2} - \mathbf{I}_m\| \leq 0.5, 1 \leq j \leq d\}.$$

We will later give a sufficient condition under which $P(\Omega_0) \rightarrow 0$, which means that, with probability approaching one, all $\hat{\Sigma}_j$ are “well behaved” in the sense that they are not too much deviated from their population values.

Define the set

$$\mathbb{C} := \{\boldsymbol{\alpha} \in \mathbb{R}^{dm} : \sum_{j \in (T^*)^c} \|\boldsymbol{\alpha}_{G_j}\|_E \leq 21 \sum_{j \in T^*} \|\boldsymbol{\alpha}_{G_j}\|_E\}.$$

The set \mathbb{C} is a cone, i.e., for any $\boldsymbol{\alpha} \in \mathbb{C}$ and $c > 0$, $c\boldsymbol{\alpha} \in \mathbb{C}$. It consists of vectors $\boldsymbol{\alpha} \in \mathbb{R}^{dm}$ such that the coordinates of $\boldsymbol{\alpha}$ in the set T^* are dominant. Such cones of dominant coordinates play an important role in the analysis of penalization methods for high dimensional statistical models. Define the \mathbb{C} -restricted eigenvalue of $\hat{\Sigma}^{1/2}$ by

$$\hat{\kappa} := \min_{\boldsymbol{\alpha} \in \mathbb{S}^{dm-1} \cap \mathbb{C}} \|\hat{\Sigma}^{1/2} \boldsymbol{\alpha}\|_E.$$

Restricted eigenvalues are originally introduced by Bickel et al. (2009) for the Lasso formulation. While the minimum eigenvalue of $\hat{\Sigma}$ is always zero when $dm \geq n$, $\hat{\kappa}$ can be positive with a high probability as long as the corresponding restricted eigenvalue of the population matrix Σ is bounded away from zero (see Lemma B.4 in Appendix B).

Put $\tilde{\mathbf{x}}_{iG_j} := \hat{\Sigma}_j^{-1/2} \tilde{\mathbf{x}}_{iG_j}$, where $\hat{\Sigma}_j^{-1/2}$ is interpreted as the generalized inverse of $\hat{\Sigma}_j^{1/2}$ if it is singular. If $\hat{\Sigma}_j = \mathbf{U} \mathbf{D} \mathbf{U}'$ denotes the spectral decomposition of $\hat{\Sigma}_j$ where \mathbf{U} is an $m \times m$ orthogonal matrix and \mathbf{D} is an $m \times m$ diagonal matrix with diagonal entries $d_1 \geq \dots \geq d_l > 0 = d_{l+1} = \dots = d_m$, then $\hat{\Sigma}_j^{-1/2}$ is given by $\hat{\Sigma}_j^{-1/2} = \mathbf{U} \text{diag}\{d_1^{-1/2}, \dots, d_l^{-1/2}, 0, \dots, 0\} \mathbf{U}'$. Invoke that on the event Ω_0 , all $\hat{\Sigma}_j$ are nonsingular. Define the random variable

$$\Lambda := \max_{1 \leq j \leq d} \left\| \sum_{i=1}^n u_i \tilde{\mathbf{x}}_{iG_j} / \sqrt{m} \right\|_E.$$

This random variable plays a “threshold” value for λ_1 .

We state a preliminary bound on \hat{g} in terms of $\|\cdot\|_{2,n}$.

Proposition 4.1. *On the event $\{\lambda_1 \geq 2\Lambda\} \cap \{\hat{\kappa} > 0\} \cap \Omega_0$, we have*

$$\|g^* - \hat{g}\|_{2,n}^2 \leq 2 \inf_{g \in \tilde{\mathcal{G}}_m^{T^*}} \|g^* - g\|_{2,n}^2 + C_2 \frac{s^* m \lambda_1^2}{\hat{\kappa}^2 n^2},$$

where C_2 is a universal constant and $\tilde{\mathcal{G}}_m^{T^*} := \{g : \mathcal{Z} \rightarrow \mathbb{R} : g(\mathbf{z}) = \sum_{j \in T^*} \sum_{k=1}^m \beta_{jk}(\psi_k(z_j) - \bar{\psi}_{jk}) \mid \mathbf{z} = (z_1, \dots, z_d)'\}, \beta_{jk} \in \mathbb{R} (j \in T^*; 1 \leq k \leq m)\}$.

To state the model selection property of the group Lasso estimator, we need another concept, namely, *group sparse eigenvalues*. For any subset $T \subset \{1, \dots, d\}$, let $\mathbb{S}_T^{dm-1} := \{\boldsymbol{\alpha} \in \mathbb{R}^{dm} : \boldsymbol{\alpha}_{G_{T^c}} = \mathbf{0}\} \cap \mathbb{S}^{dm-1}$. Define the s -th group sparse maximum eigenvalue of $\hat{\Sigma}^{1/2}$ by

$$\hat{\phi}_{\max}(s) := \max_{|T| \leq s, \boldsymbol{\alpha} \in \mathbb{S}_T^{dm-1}} \|\hat{\Sigma}^{1/2} \boldsymbol{\alpha}\|_E.$$

The next proposition gives a preliminary bound on \hat{s} , the number of components selected by the group Lasso estimator \hat{g} : $\hat{s} := |\hat{T}^0| = |\{j \in \{1, \dots, d\} : \|\hat{g}_j\|_{2,n} \neq 0\}|$.

Proposition 4.2. *Let $\hat{C} := 3n\|g^* - \hat{g}\|_{2,n}/(\sqrt{s^* m} \lambda_1)$ and $\mathcal{S} := \{s \in \{1, \dots, d\} : s > 2\hat{C}^2 \hat{\phi}_{\max}(s)^2 s^*\}$. On the event $\{\lambda_1 \geq 2\Lambda \vee 0\} \cap \Omega_0$, we have*

$$\hat{s} \leq \hat{C}^2 [\min_{s \in \mathcal{S}} \hat{\phi}_{\max}(s)^2] s^*.$$

Propositions 4.1 and 4.2 are deterministic statements, and they do not use any stochastic argument. In order to bound stochastic orders of $\|g^* - \hat{g}\|_{2,n}$ and \hat{s} , we have to determine: (i) conditions that ensure $P(\Omega_0) \rightarrow 1$; (ii) a value of λ_1 such that $\lambda_1 \geq 2\Lambda$ with probability approaching one; (iii) conditions that ensure desired asymptotic behaviors of $\hat{\kappa}$ and $\hat{\phi}_{\max}(s)$; (iv) an stochastic order of the approximation error $\inf_{g \in \tilde{\mathcal{G}}_m^{T^*}} \|g^* - g\|_{2,n}^2$. Lemmas B.1-B.5 in Appendix B are concerned with these issues. We shall comment that while the proofs of Propositions 4.1 and 4.2 are a direct

adaptation of the corresponding proofs in the Lasso case, the proofs of Lemmas B.1-B.4 are not the case because the fact that the size (m) of each group goes to infinity brings a subtle technical issue. Given Propositions 4.1 and 4.2, and Lemmas B.1-B.5 in Appendix B, we obtain the following theorem.

Theorem 4.2. *Assume conditions (C1)-(C5) and (C7). Assume further that s^*, m, d and n obey the growth condition $(s^*)^2 m \log(d \vee n)/n \rightarrow 0$, $\phi_{\max}(s) := \max_{|T| \leq s, \alpha \in \mathbb{S}_T^{dm-1}} \|\Sigma^{1/2} \alpha\|_E \lesssim 1$ for some sequence $s = s_n$ such that $s/s^* \rightarrow \infty$ and $\kappa := \min_{\alpha \in \mathbb{S}^{dm-1} \cap \mathbb{C}} \|\Sigma^{1/2} \alpha\|_E \gtrsim 1$, and $\|g^*\|_2^2 \lesssim s^*$. Take $\lambda_1 \geq A_{1,u} \sqrt{n}(1 + \sqrt{\log d/m})$ with constant $A_{1,u}$ given in Lemma B.2 in Appendix B and $m \gtrsim n^{1/(2\nu+1)}$. Then:*

$$\|g^* - \hat{g}\|_{2,n}^2 \lesssim_p \frac{s^* m \lambda_1^2}{n^2}, \quad \hat{s} \lesssim_p s^*.$$

In particular, if $m \asymp n^{1/(2\nu+1)}$ and

$$\lambda_1 \asymp \max \left\{ \sqrt{n}, \sqrt{\frac{n \log d}{m}} \right\}, \quad (4.2)$$

then we have $\|g^* - \hat{g}\|_{2,n}^2 \lesssim_p s^* \delta^2$.

Proof. See Appendix B. □

Remark 4.2. When $m \asymp n^{\nu/(2\nu+1)}$, the order of d allowed is $\log d = o\{n^{2\nu/(2\nu+1)}/(s^*)^2\}$. If $\log d \asymp n^a$ and $s^* \asymp n^b$ for some $a, b \geq 0$, the region that (a, b) is allowed is $\{(a, b) : a, b \geq 0, a + 2b < 2\nu/(2\nu+1)\}$. It is interesting to note that this region is large when ν is large, i.e., the additive components are more smooth. This indicates that the more smooth the additive components are, the larger d and s^* can be.

We consider the magnitude of missed components $\|\sum_{j \in T^* \setminus \hat{T}^0} g_j^*\|_2^2$. To this end, for a subset $T \subset \{1, \dots, d\}$, define the T -sparse minimal eigenvalue $\hat{\phi}_{\min}(T)$ of $\hat{\Sigma}$ by

$$\hat{\phi}_{\min}(T) := \min_{\alpha \in \mathbb{S}_T^{dm-1}} \|\hat{\Sigma}^{1/2} \alpha\|_E.$$

We also need a slightly stronger approximation property than condition (C7)-(c).

(C7) (c)' There exists a $g^m = \sum_{j \in T^*} g_j^m \in \mathcal{G}_m^{T^*}$ such that $\max_{T \subset T^*} \|\sum_{j \in T} (g_j - g_j^m)\|_2^2 \lesssim s^* m^{-2\nu}$.

Corollary 4.1 (Magnitude of missed components). *Assume the same conditions as in Theorem 4.2 with condition (C7)-(c) replaced by (C7)-(c)'. Assume further that $\hat{\phi}_{\min}(T^* \cup \hat{T}) \gtrsim_p 1$. Then, we have $\|\sum_{j \in T^* \setminus \hat{T}^0} g_j^*\|_2^2 \lesssim_p s^* m \lambda_1^2 / n^2$.*

This corollary clarifies sufficient conditions under which the magnitude of missed components is not larger than the bound on $\|\hat{g} - g^*\|_{2,n}^2$. When $m \asymp n^{1/(2\nu+1)}$ and λ_1 is (4.2), then, under the conditions of Corollary 4.1, $|\hat{T}^0 \setminus T^*| \lesssim_p s^*$ and $\|\sum_{j \in T^* \setminus \hat{T}^0} g_j^*\|_2^2 \lesssim_p s^* \delta^2$. In that case, the second step estimator \tilde{g} with $\hat{T} = \hat{T}^0$ and $A_{2,u,\nu} n^{-\nu/(2\nu+1)} \leq \lambda_2 \lesssim \delta$ satisfies that $\|\tilde{g} - g^*\|_2^2 + \lambda_2^2 \sum_{j \in \hat{T}} I(\tilde{g}_j) \lesssim_p s^* \delta^2$. This bound holds in general cases in which \hat{T}^0 may fail to recover T^* . If it happens that $\hat{T}^0 = T^*$ with probability approaching one, the estimator \tilde{g} (with $\lambda_2 \asymp n^{-\nu/(2\nu+1)}$) enjoys the exact oracle rate $s^* n^{-2\nu/(2\nu+1)}$. As long as taking $\lambda_2 \asymp n^{-2\nu/(2\nu+1)}$, the estimator \tilde{g} adapts to both situations.

Sufficient conditions for the perfect model selection are found in, e.g., Theorem 2 of Ravikumar et al. (2009). Unfortunately, their condition (39) does not cover our choice of the penalty level λ_1 . Note that the correspondence between their notation (left) and our notation (right) is: $p = d, d_n = m$ and $\lambda_n = \sqrt{m} \lambda_1 / n$. However, a careful inspection of their proof shows that their condition (39) can be replaced by a weaker condition that there exists some constant $C > 0$ such that

$$\frac{\lambda_n^2 n}{d_n \vee \log p} > C \text{ (in their notation), or } \frac{m \lambda_1^2}{n(m \vee \log d)} > C \text{ (in our notation),} \quad (4.3)$$

which covers our choice of the penalty level λ_1 . To see this, observe that their condition (39) is used only to ensure (85) in their appendix, which can be replaced by (in their notation) $P(\max_{j \in S^c} \|\hat{g}_j - \mu_j\| > \delta/2) \rightarrow 0$, or equivalently $P(\max_{j \in S^c} \|Z_j\| \geq \lambda_n \delta/2) \rightarrow 0$. By using first the union bound and then Theorem 7.1 of Ledoux (2001) (the Gaussian concentration inequality) similarly to the proof of our Lemma B.2 in Appendix B, it is shown that condition (4.3) is sufficient for that $P(\max_{j \in S^c} \|Z_j\| \geq \lambda_n \delta/2) \rightarrow 0$.

4.4 Comparison with other work

In this section, we briefly state connections and differences of the proposed method from some existing estimation methods for high dimensional additive models. It must be said that the literature on high dimensional additive models is now growing; so it is beyond the scope of this paper to review all the existing methods in details.

In Meier et al. (2009), the penalized least squares estimator defined by the solution to the following minimization problem is proposed:

$$\min_{c \in \mathbb{R}, g_j \in \mathcal{G}, 1 \leq j \leq d} \left[\frac{1}{2n} \sum_{i=1}^n (y_i - c - \sum_{j=1}^d g_j(z_{ij}))^2 + \sum_{j=1}^d \left\{ \tilde{\lambda}_1 \sqrt{\|g_j\|_{2,n}^2 + \tilde{\lambda}_2 I(g_j)^2} + \tilde{\lambda}_3 I(g_j)^2 \right\} \right],$$

where the term $\|g_j\|_{2,n}$ controls sparsity while the term $I(g_j)$ controls smoothness of g_j . Koltchinskii and Yuan (2010) and Raskutti et al. (2010) considered a doubly penalized

estimation method similar to Meier et al. (2009) but in a (more general) reproducing kernel Hilbert space (RKHS) formulation. Suzuki et al. (2011) further analyzed the Meier et al. (2009) method and established a faster convergence rate than Meier et al. (2009) did in a more general setting. The method proposed in this paper is thought to be a method that splits such a “double penalization” into two steps, and intends to remove a shrinkage bias caused by simultaneously penalizing sparsity and smoothness.

Huang et al. (2010) proposed a two-step estimation method different from ours. Their proposal is to construct consistent estimators of the additive components at the first step, and then to use these consistent estimators to apply the adaptive group Lasso, which is a modification of the adaptive Lasso (Zou, 2006) to the group Lasso case. In particular, they proposed to use the group Lasso for the first step estimation. To be precise, under the notation of Section 2.2, let $\hat{\beta}^0$ denote the solution to the group Lasso problem (2.3) with $\hat{\Sigma}_j^{1/2}$ replaced by \mathbf{I}_m , and use this group Lasso estimator to construct the weights: $w_j := 1/\|\hat{\beta}_{G_j}^0\|_E$ (we agree that $1/0 = \infty$). The adaptive group Lasso estimator is then defined by $\hat{g}^A(\mathbf{z}) = \sum_{j=1}^d \hat{g}_j^A(z_j)$, $\hat{g}_j^A(z_j) = \sum_{k=1}^m \hat{\beta}_{jk}^A(\psi_k(z_j) - \bar{\psi}_{jk})$, where

$$\hat{\beta}^A := \arg \min_{\beta \in \mathbb{R}^{dm}} \left[\frac{1}{2n} \sum_{i=1}^n (y_i - \tilde{\mathbf{x}}_i' \beta)^2 + \lambda_A \sum_{j=1}^d w_j \|\beta_{G_j}\|_E \right].$$

The adaptive group Lasso can be seen as a post model selection estimator. In fact, since $w_j = \infty$ when $\|\hat{\beta}_{G_j}^0\|_E = 0$, the adaptive group Lasso problem reduces to

$$\min_{\beta_{G_j}, j \in \tilde{T}} \left[\frac{1}{2n} \sum_{i=1}^n (y_i - \sum_{j \in \tilde{T}} \tilde{\mathbf{x}}_{iG_j}' \beta_{G_j})^2 + \lambda_A \sum_{j \in \tilde{T}} w_j \|\beta_{G_j}\|_E \right],$$

where $\tilde{T} := \{j \in \{1, \dots, d\} : \|\hat{\beta}_{G_j}^0\|_E \neq 0\}$. Therefore, their estimation method is similar to ours in some respect. Besides the similarity, however, there is a notable difference between two methods. Huang et al. (2010) intend to select *correctly* the nonzero additive components with probability approaching one by using the group Lasso penalty at both the first and second steps, while our method intends to ensure sparsity and smoothness of the estimator. A point to be noticed is that the analysis of Huang et al. (2010) substantially depends on the assumption that the non-zero additive components are well separated from zero in the L_2 -sense, which is, as argued in Introduction, significantly restrictive from a theoretical point of view, and it is this assumption why the adaptive group Lasso estimator can achieve the exact oracle rate in their analysis. Therefore, from a strict theoretical sense, their theoretical result is not directly comparable to ours (and Meier et al. (2009)).

5 Conclusion

In this paper we have investigated the two-step estimation of high dimensional additive models. Especially, we have derived a generic performance bound on the second step estimator, and studied the overall performance when the group Lasso is used as a first step variable selection. Diving the overall estimation procedure into two steps enables us to help shrinkage bias caused by the double penalization strategy, and we believe that the theoretical and numerical properties explored in this paper are useful suggestions to practical applications.

6 Proof of Theorem 4.1

The proof of Theorem 4.1 uses the next technical lemma. Its proof is based on a use of empirical process techniques. Define $\mu_{g_j} := \mathbb{E}[g_j(z_{1j})]$.

Lemma 6.1. *Assume conditions (C1)-(C5). Then, there exist a positive constant $C_{u,\nu}$ depending only on the distribution of u_1 and the smoothness index ν , and a positive constant $c_{q,\nu}$ depending only on c_q (given in condition (C3)) and ν such that the following holds: for any sequence of nonempty subsets $T = T_n \subset \{1, \dots, d\}$ and for any sequence of constants $\epsilon = \epsilon_n \rightarrow 0$ such that $\epsilon \geq C_{u,\nu} n^{-\nu/(2\nu+1)}$, we have, with probability approaching one:*

$$\begin{aligned}
 (i) \quad & \left| \frac{1}{n} \sum_{i=1}^n u_i g_j(z_{ij}) \right| \leq \max\{\epsilon, C_1 \sqrt{\log(s \vee n)/n}\} \sqrt{\|g_j\|_2^2 + \epsilon^2 I(g_j)^2}, \quad \forall g_j \in \mathcal{G}, \quad \forall j \in T; \\
 (ii) \quad & \left| \frac{1}{n} \sum_{i=1}^n \{g_j(z_{ij}) - \mu_{g_j}\} \right| \leq \max\{\epsilon, C_1 \sqrt{\log(s \vee n)/n}\} \sqrt{\|g_j\|_2^2 + \epsilon^2 I(g_j)^2}, \quad \forall g_j \in \mathcal{G}, \quad \forall j \in T; \\
 (iii) \quad & |||g||_{2,n}^2 - \|g\|_2^2| \leq c_{q,\nu} \epsilon^{-1/(2\nu)} \max\{\epsilon, \delta\} \left[\sum_{j=1}^d \sqrt{\|g_j\|_2^2 + \epsilon^2 I(g_j)^2} \right]^2, \quad \forall g = \sum_{j=1}^d g_j, \quad g_j \in \mathcal{G},
 \end{aligned}$$

where $s := s_n := |T|$ and $C_1 > 0$ is a universal constant.

Proof of Lemma 6.1. See Section Appendix A. □

Proof of Theorem 4.1. We first point out that because of the restriction $\sum_{i=1}^n g_j(z_{ij}) = 0$, by a standard argument, we may assume that $\hat{c} = c^* = \mathbb{E}[y_1] = 0$ for the analysis of \tilde{g} .

Let $C_{u,\nu}, c_{q,\nu}$ and C_1 be the constants given in Lemma 6.1. Take $\epsilon = \epsilon_n =$

$C_{u,\nu}n^{-\nu/(2\nu+1)}$ and $\lambda_2 \geq \sqrt{2}\epsilon$. Define the events

$$\begin{aligned}\Omega_1 &:= \text{event (i) of Lemma 6.1 with } T = T^*, \\ \Omega_2 &:= \text{event (i) of Lemma 6.1 with } T = (T^*)^c, \\ \Omega_3 &:= \text{event (ii) of Lemma 6.1 with } T = T^*, \\ \Omega_4 &:= \text{event (ii) of Lemma 6.1 with } T = (T^*)^c, \\ \Omega_5 &:= \text{event (iii) of Lemma 6.1.}\end{aligned}$$

In what follows, we go through the proof on the events $\cap_{k=1}^5 \Omega_k$. Note that the probability of this event goes to one. Because $s^* = |T^*| \leq n$, we may assume that $C_1\sqrt{\log(s^* \vee n)/n} = C_1\sqrt{\log n/n} \leq \epsilon$ in events (i) and (ii) of Lemma 6.1 with $T = T^*$. Let $\varrho := \varrho_n := \max\{\epsilon, C_1\sqrt{\log d/n}\}$. Invoke that $\varrho \asymp \delta$. We may assume that $\varrho \leq 1$. For $j \notin \hat{T}$, we agree that $\tilde{g}_j \equiv 0$.

Because of the optimality of \tilde{g} ,

$$\frac{1}{2n} \sum_{i=1}^n (y_i - \sum_{j \in \hat{T}} \tilde{g}_j(z_{ij}))^2 + \lambda_2^2 \sum_{j \in \hat{T}} I(\tilde{g}_j)^2 \leq \frac{1}{2n} \sum_{i=1}^n (y_i - \sum_{j \in \hat{T}} g_j^*(z_{ij}))^2 + \lambda_2^2 \sum_{j \in \hat{T}} I(g_j^*)^2.$$

Then, using the relation

$$(y_i - g(z_i))^2 = u_i^2 + 2u_i(g^*(z_i) - g(z_i)) + (g^*(z_i) - g(z_i))^2,$$

we have (note that $\tilde{g}_j \equiv 0$ for $j \notin \hat{T}$)

$$\begin{aligned}& \frac{1}{2} \|g^* - \tilde{g}\|_{2,n}^2 + \lambda_2^2 \sum_{j \in \hat{T}} I(\tilde{g}_j)^2 \\ & \leq \sum_{j \in \hat{T}} \left[\frac{1}{n} \sum_{i=1}^n u_i \{\tilde{g}_j(z_{ij}) - g_j^*(z_{ij})\} \right] + \lambda_2^2 \sum_{j \in \hat{T}} I(g_j^*)^2 + \frac{1}{2} \|\sum_{j \in T^* \setminus \hat{T}} g_j^*\|_{2,n}^2.\end{aligned}$$

Using the facts that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, $ab \leq 0.5(a^2 + b^2)$ and $(a+b)^2 \leq 2(a^2 + b^2)$,

$$\begin{aligned}& \sum_{j \in T^* \cap \hat{T}} \left[\frac{1}{n} \sum_{i=1}^n u_i \{\tilde{g}_j(z_{ij}) - g_j^*(z_{ij})\} \right] \\ & \leq \epsilon \sum_{j \in T^* \cap \hat{T}} \sqrt{\|\tilde{g}_j - g_j^*\|_2^2 + \epsilon^2 I(\tilde{g}_j - g_j^*)^2} \quad (\because \Omega_1) \\ & \leq \epsilon \sum_{j \in T^* \cap \hat{T}} \|\tilde{g}_j - g_j^*\|_2 + \epsilon^2 \sum_{j \in T^* \cap \hat{T}} I(\tilde{g}_j - g_j^*) \\ & \leq \epsilon \sqrt{|T^* \cap \hat{T}| \sum_{j \in T^* \cap \hat{T}} \|\tilde{g}_j - g_j^*\|_2^2 + 0.5\epsilon^2 |T^* \cap \hat{T}| + 0.5\epsilon^2 \sum_{j \in T^* \cap \hat{T}} I(\tilde{g}_j - g_j^*)^2} \\ & \leq \epsilon \sqrt{|T^* \cap \hat{T}| \sum_{j \in T^* \cap \hat{T}} \|\tilde{g}_j - g_j^*\|_2^2 + 0.5\epsilon^2 |T^* \cap \hat{T}| + \epsilon^2 \sum_{j \in T^* \cap \hat{T}} I(\tilde{g}_j)^2 + \epsilon^2 \sum_{j \in T^* \cap \hat{T}} I(g_j^*)^2}.\end{aligned}$$

For any fixed $b > 0$,

$$\begin{aligned} \epsilon \sqrt{|T^* \cap \hat{T}| \sum_{j \in T^* \cap \hat{T}} \|\tilde{g}_j - g_j^*\|_2^2} &= \sqrt{2b\epsilon^2 |T^* \cap \hat{T}| \times \frac{1}{2b} \sum_{j \in T^* \cap \hat{T}} \|\tilde{g}_j - g_j^*\|_2^2} \\ &\leq b\epsilon^2 |T^* \cap \hat{T}| + \frac{1}{4b} \sum_{j \in T^* \cap \hat{T}} \|\tilde{g}_j - g_j^*\|_2^2. \end{aligned}$$

Similarly, we have

$$\begin{aligned} &\sum_{j \in \hat{T} \setminus T^*} \left[\frac{1}{n} \sum_{i=1}^n u_i \{\tilde{g}_j(z_{ij}) - g_j^*(z_{ij})\} \right] \\ &= \sum_{j \in \hat{T} \setminus T^*} \left[\frac{1}{n} \sum_{i=1}^n u_i \tilde{g}_j(z_{ij}) \right] \\ &\leq \varrho \sum_{j \in \hat{T} \setminus T^*} \sqrt{\|\tilde{g}_j\|_2^2 + \epsilon^2 I(\tilde{g}_j)^2} \quad (\because \Omega_2) \\ &\leq \varrho \sum_{j \in \hat{T} \setminus T^*} \|\tilde{g}_j\|_2 + \varrho\epsilon \sum_{j \in \hat{T} \setminus T^*} I(\tilde{g}_j) \\ &\leq \varrho \sqrt{|\hat{T} \setminus T^*| \sum_{j \in \hat{T} \setminus T^*} \|\tilde{g}_j\|_2^2 + 0.5\varrho^2 |\hat{T} \setminus T^*| + 0.5\epsilon^2 \sum_{j \in \hat{T} \setminus T^*} I(\tilde{g}_j)^2} \\ &\leq \frac{1}{4b} \sum_{j \in \hat{T} \setminus T^*} \|\tilde{g}_j\|_2^2 + (b + 0.5)\varrho^2 |\hat{T} \setminus T^*| + 0.5\epsilon^2 \sum_{j \in \hat{T} \setminus T^*} I(\tilde{g}_j)^2. \end{aligned}$$

Thus, we have

$$\begin{aligned} \frac{1}{2} \|g^* - \tilde{g}\|_{2,n}^2 + (\lambda_2^2 - \epsilon^2) \sum_{j \in \hat{T}} I(\tilde{g}_j)^2 &\leq \frac{1}{4b} \sum_{j \in \hat{T}} \|\tilde{g}_j - g_j^*\|_2^2 + (b + 0.5)(\epsilon^2 |T^* \cap \hat{T}| + \varrho^2 |\hat{T} \setminus T^*|) \\ &\quad + (\lambda_2^2 + \epsilon^2) \sum_{j \in T^* \cap \hat{T}} I(g_j^*)^2 + \frac{1}{2} \|\sum_{j \in T^* \setminus \hat{T}} g_j^*\|_{2,n}^2. \quad (6.1) \end{aligned}$$

Recall the definition of $\alpha(T)$. Invoke now that

$$\begin{aligned} \sum_{j \in \hat{T}} \|\tilde{g}_j - g_j^*\|_2^2 &\leq \sum_{j \in T^* \cup \hat{T}} \|\tilde{g}_j - g_j^*\|_2^2 \\ &\leq 2 \sum_{j \in T^* \cup \hat{T}} \|(\tilde{g}_j - \mu_{\tilde{g}_j}) - g_j^*\|_2^2 + 2 \sum_{j \in \hat{T}} \mu_{\tilde{g}_j}^2 \quad (\mu_{g_j} := \mathbb{E}[g_j(z_{1j})]) \\ &\leq 2\alpha(T^* \cup \hat{T}) \|(\tilde{g} - \mu_{\tilde{g}}) - g^*\|_2^2 + 2 \sum_{j \in \hat{T}} \mu_{\tilde{g}_j}^2 \\ &\leq 2\alpha(T^* \cup \hat{T}) \|\tilde{g} - g^*\|_2^2 + 2 \sum_{j \in \hat{T}} \mu_{\tilde{g}_j}^2. \end{aligned}$$

Because of the restriction $\sum_{i=1}^n \tilde{g}(z_{ij}) = 0$, $\mu_{\tilde{g}_j} = -n^{-1} \sum_{i=1}^n (\tilde{g}_j(z_{ij}) - \mu_{\tilde{g}_j})$, so that, because of the event Ω_3 , for all $j \in T^* \cap \hat{T}$,

$$\begin{aligned} \mu_{\tilde{g}_j}^2 &\leq \epsilon^2 \|\tilde{g}_j\|_2^2 + \epsilon^4 I(\tilde{g}_j)^2 \\ &\leq 2\epsilon^2 \|\tilde{g}_j - g_j^*\|_2^2 + 2\epsilon^2 \|g_j^*\|_2^2 + \epsilon^4 I(\tilde{g}_j)^2, \end{aligned}$$

while because of the event Ω_4 , for all $j \in \hat{T} \setminus T^*$, $\mu_{\tilde{g}_j}^2 \leq \varrho^2 \|\tilde{g}_j\|_2^2 + \varrho^2 \epsilon^2 I(\tilde{g}_j)^2$. Thus, noting that $\varrho \geq \epsilon$,

$$(1 - \max\{4\epsilon^2, 2\varrho^2\}) \sum_{j \in \hat{T}} \|\tilde{g}_j - g_j^*\|_2^2 \leq 2\alpha(T^* \cup \hat{T}) \|\tilde{g} - g^*\|_2^2 + 4\epsilon^2 \sum_{j \in T^* \cap \hat{T}} \|g_j^*\|_2^2 + 2\varrho^2 \epsilon^2 \sum_{j \in \hat{T}} I(\tilde{g}_j)^2,$$

so that for n large enough (such that $\max\{4\epsilon^2, 2\varrho^2\} \leq 0.5$),

$$\sum_{j \in \hat{T}} \|\tilde{g}_j - g_j^*\|_2^2 \leq 4\alpha(T^* \cup \hat{T}) \|\tilde{g} - g^*\|_2^2 + 8\epsilon^2 \sum_{j \in T^* \cap \hat{T}} \|g_j^*\|_2^2 + 4\varrho^2 \epsilon^2 \sum_{j \in \hat{T}} I(\tilde{g}_j)^2. \quad (6.2)$$

Substituting (6.2) into (6.1), we obtain

$$\begin{aligned} &\frac{1}{2} \|g^* - \tilde{g}\|_{2,n}^2 + \left(\lambda_2^2 - \epsilon^2 - \frac{\varrho^2 \epsilon^2}{b} \right) \sum_{j \in \hat{T}} I(\tilde{g}_j)^2 \\ &\leq \frac{\alpha(T^* \cup \hat{T})}{b} \|\tilde{g} - g^*\|_2^2 + (b + 0.5)(\epsilon^2 |T^* \cap \hat{T}| + \varrho^2 |\hat{T} \setminus T^*|) + \frac{2\epsilon^2}{b} \sum_{j \in T^* \cap \hat{T}} \|g_j^*\|_2^2 \\ &\quad + (\lambda_2^2 + \epsilon^2) \sum_{j \in T^* \cap \hat{T}} I(g_j^*)^2 + \frac{1}{2} \|\sum_{j \in T^* \setminus \hat{T}} g_j^*\|_{2,n}^2. \end{aligned} \quad (6.3)$$

We next consider a lower bound on $\|g^* - \tilde{g}\|_{2,n}^2$. Observe that

$$\begin{aligned} &\|g^* - \tilde{g}\|_{2,n}^2 \\ &\geq \|g^* - \tilde{g}\|_2^2 - c_{q,\nu} \epsilon^{-1/(2\nu)} \max\{\epsilon, \delta\} \left[\sum_{j \in T^* \cup \hat{T}} \sqrt{\|g_j^* - \tilde{g}_j\|_2^2 + \epsilon^2 I(g_j^* - \tilde{g}_j)^2} \right]^2 \quad (\because \Omega_5) \\ &\geq \|g^* - \tilde{g}\|_2^2 - c_{q,\nu} \epsilon^{-1/(2\nu)} \max\{\epsilon, \delta\} |T^* \cup \hat{T}| \sum_{j \in T^* \cup \hat{T}} \{\|g_j^* - \tilde{g}_j\|_2^2 + \epsilon^2 I(g_j^* - \tilde{g}_j)^2\} \\ &\geq \|g^* - \tilde{g}\|_2^2 - c_{q,\nu} \epsilon^{-1/(2\nu)} \max\{\epsilon, \delta\} |T^* \cup \hat{T}| \left\{ \alpha(T^* \cup \hat{T}) \|g^* - \tilde{g}\|_2^2 + \epsilon^2 \sum_{j \in T^* \cup \hat{T}} I(g_j^* - \tilde{g}_j)^2 \right\} \\ &\geq (1 - \hat{c}_1) \|g^* - \tilde{g}\|_2^2 - 2\hat{c}_2 \epsilon^2 \sum_{j \in T^*} I(g_j^*)^2 - 2\hat{c}_2 \epsilon^2 \sum_{j \in \hat{T}} I(\tilde{g}_j)^2, \end{aligned}$$

where $\hat{c}_1 := c_{q,\nu} \epsilon^{-1/(2\nu)} \max\{\epsilon, \delta\} |T^* \cup \hat{T}| \alpha(T^* \cup \hat{T})$ and $\hat{c}_2 := c_{q,\nu} \epsilon^{-1/(2\nu)} \max\{\epsilon, \delta\} |T^* \cup \hat{T}|$

\hat{T} . Substituting this inequality to (6.3), we have

$$\begin{aligned}
& \left(\frac{1}{2} - \frac{\hat{c}_1}{2} - \frac{\alpha(T^* \cup \hat{T})}{b} \right) \|g^* - \tilde{g}\|_2^2 + \left(\lambda_2^2 - \epsilon^2 - \frac{\varrho^2 \epsilon^2}{b} - \hat{c}_2 \epsilon^2 \right) \sum_{j \in \hat{T}} I(\tilde{g}_j)^2 \\
& \leq (b + 0.5)(\epsilon^2 |T^* \cap \hat{T}| + \varrho^2 |\hat{T} \setminus T^*|) + \frac{2\epsilon^2}{b} \sum_{j \in T^* \cap \hat{T}} \|g_j^*\|_2^2 + (\lambda_2^2 + \epsilon^2 + \hat{c}_2 \epsilon^2) \sum_{j \in T^* \cap \hat{T}} I(g_j^*)^2 \\
& \quad + \frac{1}{2} \|\sum_{j \in T^* \setminus \hat{T}} g_j^*\|_{2,n}^2 + \hat{c}_2 \epsilon^2 \sum_{j \in T^* \setminus \hat{T}} I(g_j^*)^2.
\end{aligned}$$

We wish to bound the term $\|\sum_{j \in T^* \setminus \hat{T}} g_j^*\|_{2,n}^2$. Observe that

$$\begin{aligned}
& \|\sum_{j \in T^* \setminus \hat{T}} g_j^*\|_{2,n}^2 \\
& \leq \|\sum_{j \in T^* \setminus \hat{T}} \hat{T} g_j^*\|_2^2 + c_{q,\nu} \epsilon^{-1/(2\nu)} \max\{\epsilon, \delta\} \left[\sum_{j \in T^* \setminus \hat{T}} \sqrt{\|g_j^*\|_2^2 + \epsilon^2 I(g_j^*)^2} \right]^2 \quad (\cdot \cdot \Omega_5) \\
& \leq \|\sum_{j \in T^* \setminus \hat{T}} \hat{T} g_j^*\|_2^2 + c_{q,\nu} \epsilon^{-1/(2\nu)} \max\{\epsilon, \delta\} |T^* \setminus \hat{T}| \sum_{j \in T^* \setminus \hat{T}} \{\|g_j^*\|_2^2 + \epsilon^2 I(g_j^*)^2\} \\
& \leq \|\sum_{j \in T^* \setminus \hat{T}} \hat{T} g_j^*\|_2^2 + c_{q,\nu} \epsilon^{-1/(2\nu)} \max\{\epsilon, \delta\} |T^* \setminus \hat{T}| \left\{ \alpha(T^* \setminus \hat{T}) \|\sum_{j \in T^* \setminus \hat{T}} g_j^*\|_2^2 + \epsilon^2 \sum_{j \in T^* \setminus \hat{T}} I(g_j^*)^2 \right\} \\
& \leq (1 + \hat{c}_3) \|\sum_{j \in T^* \setminus \hat{T}} \hat{T} g_j^*\|_2^2 + \hat{c}_4 \epsilon^2 \sum_{j \in T^*} I(g_j^*)^2,
\end{aligned}$$

where $\hat{c}_3 := c_{q,\nu} \epsilon^{-1/(2\nu)} \max\{\epsilon, \delta\} |T^* \setminus \hat{T}| \alpha(T^* \setminus \hat{T})$ and $\hat{c}_4 := c_{q,\nu} \epsilon^{-1/(2\nu)} \max\{\epsilon, \delta\} |T^* \setminus \hat{T}|$.

Therefore, we have

$$\begin{aligned}
& \left(\frac{1}{2} - \frac{\hat{c}_1}{2} - \frac{\alpha(T^* \cup \hat{T})}{b} \right) \|g^* - \tilde{g}\|_2^2 + \left(\lambda_2^2 - \epsilon^2 - \frac{\varrho^2 \epsilon^2}{b} - \hat{c}_2 \epsilon^2 \right) \sum_{j \in \hat{T}} I(\tilde{g}_j)^2 \\
& \leq (b + 0.5)(\epsilon^2 |T^* \cap \hat{T}| + \varrho^2 |\hat{T} \setminus T^*|) + \frac{2\epsilon^2}{b} \sum_{j \in T^* \cap \hat{T}} \|g_j^*\|_2^2 + (\lambda_2^2 + \epsilon^2 + \hat{c}_2 \epsilon^2) \sum_{j \in T^* \cap \hat{T}} I(g_j^*)^2 \\
& \quad + \frac{(1 + \hat{c}_3)}{2} \|\sum_{j \in T^* \setminus \hat{T}} g_j^*\|_2^2 + \left(\hat{c}_2 + \frac{\hat{c}_4}{2} \right) \epsilon^2 \sum_{j \in T^* \setminus \hat{T}} I(g_j^*)^2.
\end{aligned}$$

Taking $b = 4\alpha(T^* \cup \hat{T}) \geq 4$ and noting that $\varrho \leq 1$, we have

$$\begin{aligned}
& \left(\frac{1}{4} - \frac{\hat{c}_1}{2} \right) \|g^* - \tilde{g}\|_2^2 + \left\{ \lambda_2^2 - \left(\frac{5}{4} + \hat{c}_2 \right) \epsilon^2 \right\} \sum_{j \in \hat{T}} I(\tilde{g}_j)^2 \\
& \leq \{4\alpha(T^* \cup \hat{T}) + 0.5\}(\epsilon^2 |T^* \cap \hat{T}| + \varrho^2 |\hat{T} \setminus T^*|) + \frac{\epsilon^2}{2} \|g^*\|_2^2 + (\lambda_2^2 + \epsilon^2 + \hat{c}_2 \epsilon^2) \sum_{j \in T^* \cap \hat{T}} I(g_j^*)^2 \\
& \quad + \frac{(1 + \hat{c}_3)}{2} \|\sum_{j \in T^* \setminus \hat{T}} g_j^*\|_2^2 + \left(\hat{c}_2 + \frac{\hat{c}_4}{2} \right) \epsilon^2 \sum_{j \in T^* \setminus \hat{T}} I(g_j^*)^2,
\end{aligned}$$

where we have used the inequality

$$\frac{2}{b} \sum_{j \in T^* \cap \hat{T}} \|g_j^*\|_2^2 \leq \frac{2}{b} \sum_{j \in T^*} \|g_j^*\|_2^2 \leq \frac{2\alpha(T^*)}{b} \|g^*\|_2^2 \leq 0.5 \|g^*\|_2^2.$$

Because $\hat{c}_1 = o_p(1)$, $\hat{c}_2 = o_p(1)$, $\hat{c}_3 = o_p(1)$ and $\hat{c}_4 = o_p(1)$ by condition (C6), we have $\hat{c}_1 \leq 1/4$, $\hat{c}_2 \leq 1/2$, $\hat{c}_3 \leq 1$ and $\hat{c}_4 \leq 1$ with probability approaching one. Define the event

$$\Omega_6 := \{\hat{c}_1 \leq 1/4, \hat{c}_2 \leq 1/2, \hat{c}_3 \leq 1, \hat{c}_4 \leq 1\}.$$

Recall that $\lambda_2^2 \geq 2\epsilon^2$. Therefore, on the event $\cap_{k=1}^6 \Omega_k$, we have

$$\begin{aligned} \frac{1}{8} \|g^* - \tilde{g}\|_2^2 + \frac{\lambda_2^2}{4} \sum_{j \in \hat{T}} I(\tilde{g}_j)^2 &\leq \{4\alpha(T^* \cup \hat{T}) + 0.5\}(\epsilon^2 |T^* \cap \hat{T}| + \varrho^2 |\hat{T} \setminus T^*|) \\ &+ \frac{\epsilon^2}{2} \|g^*\|_2^2 + (\lambda_2^2 + 1.5\epsilon^2) \sum_{j \in T^* \cap \hat{T}} I(g_j^*)^2 + \|\sum_{j \in T^* \setminus \hat{T}} g_j^*\|_2^2 + \epsilon^2 \sum_{j \in T^* \setminus \hat{T}} I(g_j^*)^2. \end{aligned}$$

The desired conclusion follows from the fact that $P(\cap_{k=1}^6 \Omega_k) \rightarrow 1$. \square

Acknowledgments

The author acknowledges Dr. Lukas Meier for sharing his codes used in Meier et al. (2009) and Dr. Isamu Nagai for helping the numerical experiments. Most of the work was done when the author was visiting Department of Economics, MIT. He greatly acknowledges their hospitality. This work was supported by the Grant-in-Aid for Young Scientists (B) (22730179) from the JSPS.

References

- Bach, F.R. (2008). Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.* **9** 1179-1225.
- Belloni, A. and Chernozhukov, V. (2011a). ℓ_1 -penalized quantile regression for high dimensional sparse models. *Ann. Statist.* **39** 82-130.
- Belloni, A. and Chernozhukov, V. (2011b). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, to appear.
- Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705-1732.

- Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.
- Bunea, F., Tsybakov, A. and Wegkamp, M. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35** 1674-1697.
- Bunea, F., Tsybakov, A. and Wegkamp, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169-184.
- Candès, E.J. and Plan, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *Ann. Statist.* **37** 2145-2177.
- Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models. *J. Amer. Stat. Assoc.* **106** 544-557.
- Gabushin, V.N. (1967). Inequalities for the norms of a function and its derivatives in metric L_p . *Mathematicheskije Zametki* **1** 291-298.
- Huang, J., Horowitz, J.L. and Wei, F. (2010). Variable selection in nonparametric additive models. *Ann. Statist.* **38** 2282-2313.
- Huang, J. and Zhang, T. (2010). The benefit of group sparsity. *Ann. Statist.* **38** 1978-2004.
- Kato, K. (2011). Group Lasso for high dimensional sparse quantile regression models. Preprint.
- Koltchinskii, V. and Yuan, M. (2010). Sparsity in multiple kernel learning. *Ann. Statist.* **38** 3660-3695.
- Lin, Y. and Zhang, H.H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34** 2272-2297.
- Luinici, K., Pontil, M., Tsybakov, A.B. and van de Geer, S.A. (2010). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164-2204.
- Ledoux, M. (2001). *The Concentration of Measure Phenomenon*. American Mathematical Society.
- Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces*. Springer-Verlag.
- Massart, P. (2000). About the constants in Talagrand's concentration inequality for empirical processes. *Ann. Probab.* **28** 863-884.

- Meier, L., van de Geer, S.A. and Bühlmann, P. (2009). High-dimensional additive modeling. *Ann. Statist.* **37** 3779-3821.
- Meinshausen, N. and Yu, B. (2009). Lasso-type recovery and sparse representations for high-dimensional data. *Ann. Statist.* **37** 246-270.
- Mendelson, S. and Tomczak-Jaegermann, N. (2008). Suggaussian embedding theorem. *Israel J. Math.* **164** 349-364.
- Nardi, Y. and Rinaldo, A. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electron. J. Stat.* **2** 605-633.
- Negahban, S., Ravikumar, P., Wainwright, M.J. and Yu, B. (2010). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. Preprint.
- Newey, W.K. (1997). Convergence rates and asymptotic normality for series estimators. *J. Econometrics* **79** 147-168.
- Raskutti, G., Wainwright, M.J. and Yu, B. (2010). Minmax-optimal rates for sparse additive models over kernel classes via convex programming. Preprint.
- Ravikumar, P., Liu, H., Lafferty, J. and Wasserman, L. (2009). Sparse additive models. *J.R. Stat. Soc. Ser. B Stat. Methodol.* **71** 1009-1030.
- Rudelson, M. (1999). Random vectors in the isotropic position. *J. Funct. Anal.* **164** 60-72.
- Schumaker, L.L. (2007). *Spline Functions: Basic Theory, 3rd edition*. Cambridge University Press.
- Suzuki, T., Tomioka, R. and Sugiyama, M. (2011). Fast convergence rate of multiple kernel learning with elastic-net regularization. Preprint.
- Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126** 503-563.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J.R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267-288.
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J.R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49-67.

- van de Geer, S.A. (2000). *Empirical Processes in M-estimation*. Cambridge University Press.
- van de Geer, S.A. (2008). High-dimensional generalized linear models and the Lasso. *Ann. Statist.* **36** 614-645.
- van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag.
- Wainwright, M.J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using L_1 -constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183-2202.
- Wei, F. and Huang, J. (2010). Consistent group selection in high-dimensional linear regression. *Bernoulli* **16** 1369-1384.
- Wood, S.N. (2006). *Generalized Additive Models: an Introduction with R*. Chapman and Hall.
- Zhao, P. and Yu, B. (2007). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541-2567.
- Zhang, T. (2009). Some sharp performance bounds for least squares regression with L_1 penalization. *Ann. Statist.* **37** 2109-2144.
- Zhang, C.H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567-1594.
- Zhou, S. (2009). Restricted eigenvalue conditions on subgaussian random matrices. Preprint.
- Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Stat. Assoc.* **101** 1418-1429.

A Proof of Lemma 6.1

In the proofs below, we agree that C denotes a universal constant, and its value may change from line to line. The same rule applies to Appendix B.

A.1 Preliminary lemmas

In this section, we collect some preliminary results used in the proof of Lemma 6.1. We begin with introducing an interpolation inequality by Gabushin (1967).

Lemma A.1 (Gabushin (1967)). *For any $f \in W_2^\nu([0, 1])$ with positive integer ν ,*

$$\|f\|_\infty \leq \begin{cases} K \|f\|_{L_2(\lambda)}^{(2\nu-1)/(2\nu)} \|f^{(\nu)}\|_{L_2(\lambda)}^{1/(2\nu)}, & \text{if } \|f^{(\nu)}\|_{L_2(\lambda)} \neq 0, \\ K \|f\|_{L_2(\lambda)}, & \text{if } \|f^{(\nu)}\|_{L_2(\lambda)} = 0, \end{cases}$$

where K is a constant independent of f , and $\|\cdot\|_{L_2(\lambda)}$ denotes the L_2 -norm with respect to the Lebesgue measure λ on $[0, 1]$.

We also use the next lemma. For any probability measure Q on $[0, 1]$, define the $\nu \times \nu$ matrix

$$\Sigma_{Q,\nu} := \mathbb{E}_{Z \sim Q} \left[(1, Z, \dots, Z^{\nu-1}) \begin{pmatrix} 1 \\ Z \\ \vdots \\ Z^{\nu-1} \end{pmatrix} \right].$$

Lemma A.2. *Let Q be any probability measure on $[0, 1]$ such that the matrix $\Sigma_{Q,\nu}$ is non-singular. For any $f \in W_2^\nu([0, 1])$ (ν is a positive integer), there exist functions $f^{[1]}$ and $f^{[2]}$ such that (i) $f = f^{[1]} + f^{[2]}$; (ii) $\|f^{[1]}\|_\infty \leq \text{const.} \times I(f^{[1]})$ (the constant is independent of f); (iii) $f^{[2]}$ is a polynomial function on $[0, 1]$ of degree $\nu - 1$; and (iv) $\int f^{[1]} f^{[2]} dQ = 0$.*

Proof (sketch). Lemma A.2 is used in Meier et al. (2009) but without proof. For the sake of completeness, we provide a sketch of the proof. Take any $f \in W_2^\nu([0, 1])$. It is standard to see that there exist functions $f^{[1]}$ and $f^{[2]}$ such that $f = f^{[1]} + f^{[2]}$, $\|f^{[1]}\|_\infty \leq I(f^{[1]})$ and $f^{[2]}$ is a polynomial function on $[0, 1]$ of degree $\nu - 1$ (use Taylor's theorem). Let $\tilde{f}^{[1]}$ denote the orthogonal projection (in $L_2(Q)$) of $f^{[1]}$ onto the space of all polynomial functions of degree $\nu - 1$. Then, using the fact that $\Sigma_{Q,\nu}$ is non-singular, by a simple algebra, each coefficient of z^k ($k = 0, \dots, \nu - 1$) in $\tilde{f}^{[1]}$ is bounded by $K \|f^{[1]}\|_\infty \leq K I(f^{[1]})$, so that $\|\tilde{f}^{[1]}\|_\infty \leq K' I(f^{[1]})$ (K and K' are constants independent of f). Replacing $f^{[1]}$ by $f^{[1]} - \tilde{f}^{[1]}$ and $f^{[2]}$ by $f^{[2]} + \tilde{f}^{[1]}$, we obtain the desired conclusion. \square

It is standard to see that $\Sigma_{Q,\nu}$ is non-singular if the density of Q is bounded away from zero on $[0, 1]$. The next lemma is due to Corollary 5 of Meier et al. (2009), which is basically deduced from an entropy integral argument and a peeling argument

(such techniques are described in Chapter 8 of van de Geer (2000)). Recall that $I(f)^2 := \int_0^1 f^{(\nu)}(z)^2 dz$.

Lemma A.3 (Essentially Meier et al. (2009), Corollary 5). *Let ξ_1, \dots, ξ_n be i.i.d. from a distribution Q on $[0, 1]$ such that $\Sigma_{Q, \nu}$ is non-singular, and let $\sigma_1, \dots, \sigma_n$ be independent Rademacher random variables independent of ξ_1, \dots, ξ_n . Let $E_\sigma[\cdot]$ denote the conditional expectation with respect to $\sigma_1, \dots, \sigma_n$ given ξ_1, \dots, ξ_n . Then, there exists a positive constant C_ν depending only on ν such that for all $\epsilon \geq C_\nu n^{-\nu/(2\nu+1)}$,*

$$E \left[\sup_{f \in W_2^\nu([0,1])} \frac{|n^{-1} \sum_{i=1}^n \sigma_i f(\xi_i)|}{\sqrt{\|f\|_2^2 + \epsilon^2 I(f)^2}} \right] \leq C_\nu \epsilon, \quad E_\sigma \left[\sup_{\substack{f \in W_2^\nu([0,1]) \\ I(f) \leq 1, \|f\|_\infty \leq 1}} \frac{|n^{-1} \sum_{i=1}^n \sigma_i f(\xi_i)|}{\sqrt{\|f\|_{2,n}^2 + \epsilon^2 I(f)^2}} \right] \leq C_\nu \epsilon,$$

where $\|f\|_{2,n}^2 := n^{-1} \sum_{i=1}^n f(\xi_i)^2$ and $\|f\|_2^2 := E[f(\xi_1)^2]$. The conclusion is true when $\sigma_1, \dots, \sigma_n$ are independent standard normal.

Proof. The first inequality is Corollary 5 of Meier et al. (2009). Note that their s, α and γ correspond to $s = \nu, \alpha = 1 - 1/(2\nu)$ and $\gamma = 2/(2\nu + 1)$ in our notation. The second inequality can be shown in a similar way. \square

Addendum A.1. It is clear that, under the same conditions of Lemma A.2, for any constant $K > 0$,

$$E_\sigma \left[\sup_{\substack{f \in W_2^\nu([0,1]) \\ I(f) \leq 1, \|f\|_\infty \leq K}} \frac{|n^{-1} \sum_{i=1}^n \sigma_i f(\xi_i)|}{\sqrt{\|f\|_{2,n}^2 + \epsilon^2 I(f)^2}} \right] \leq C_\nu \epsilon.$$

The next two lemmas compare the empirical and population L_2 -norms on the class \mathcal{G} uniformly over the distributions of z_{1j} ($j \in T$).

Lemma A.4. *Assume conditions (C1), (C3) and (C4). Let T be any subset of $\{1, \dots, d\}$ and $s := |T|$. Let C_ν be the constant given in Lemma A.1. Then, there exists a positive constant $C_{q,\nu}$ depending only on c_q (which is given in condition (C3)) and ν such that, as long as*

$$C_\nu n^{-\nu/(2\nu+1)} \leq \epsilon \leq 1 \text{ and } C_{q,\nu} \epsilon^{-1/(2\nu)} \max\{\epsilon, \sqrt{\log(s \vee n)/n}\} \leq 0.5,$$

with probability at least $1 - (s \vee n)^{-1}$,

$$\|g_j\|_{2,n}^2 \leq 1.5 \|g_j\|_2^2 + 0.5 \epsilon^2 I(g_j)^2, \quad \forall g_j \in \mathcal{G}, \quad \forall j \in T.$$

Proof. In this proof, $C_{q,\nu}$ denotes some positive constant depending only on c_q and ν . Its value may change from line to line. Pick any $j \in T$. For a constant $\epsilon \in (0, 1]$ specified later, define

$$Z_j := \sup_{g_j \in \mathcal{G}} \frac{\|g_j\|_{2,n}^2 - \|g_j\|_2^2}{(\|g_j\|_2^2 + \epsilon^2 I(g_j)^2)}.$$

By Lemma A.1, invoke that when $I(g_j) \neq 0$,

$$\begin{aligned} \|g_j\|_\infty &\leq C_{q,\nu} \|g_j\|_2^{(2\nu-1)/(2\nu)} I(g_j)^{1/(2\nu)} \\ &\leq C_{q,\nu} \|g_j\|_2^{(2\nu-1)/(2\nu)} (\epsilon I(g_j))^{1/(2\nu)} \epsilon^{-1/(2\nu)} \\ &\leq C_{q,\nu} \epsilon^{-1/(2\nu)} \sqrt{\|g_j\|_2^2 + \epsilon^2 I(g_j)^2}, \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}[g_j(z_{1j})^4] &\leq \|g_j\|_\infty^2 \|g_j\|_2^2 \\ &\leq C_{q,\nu} \epsilon^{-1/\nu} (\|g_j\|_2^2 + \epsilon^2 I(g_j)^2)^2. \end{aligned}$$

Even if $I(g_j) = 0$, these inequalities hold (with a suitable change to the constant $C_{q,\nu}$ if necessary) since $\epsilon \leq 1$. Thus, by Massart's (2000) form of Talagrand's (1996) inequality, for all $t > 0$, with probability at least $1 - e^{-t}$,

$$Z_j \leq 2\mathbb{E}[Z_j] + C_{q,\nu} \sqrt{\epsilon^{-1/\nu} t/n} + C_{q,\nu} \epsilon^{-1/\nu} t/n.$$

Applying Lemma A.3 with the symmetrization inequality (van der Vaart and Wellner, 1996, Lemma 2.3.1) and the contraction principle (Ledoux and Talagrand, 1991, Theorem 4.12), for all $\epsilon \geq C_\nu n^{-\nu/(2\nu+1)}$, we have $\mathbb{E}[Z_j] \leq C_{q,\nu} \epsilon^{(2\nu-1)/(2\nu)}$. Therefore, taking $t = 2 \log(s \vee n)$, we have, with probability at least $1 - (s \vee n)^{-2}$,

$$Z_j \leq C_{q,\nu} \max\{\epsilon^{(2\nu-1)/(2\nu)}, \epsilon^{-1/(2\nu)} \sqrt{\log(s \vee n)/n}, \epsilon^{-1/\nu} \log(s \vee n)/n\}.$$

By the union bound, the above inequality simultaneously holds for all $j \in T$ with probability at least $1 - (s \vee n)^{-1}$. The desired conclusion now follows from the additional restriction that $C_{q,\nu} \epsilon^{-1/(2\nu)} \max\{\epsilon, \sqrt{\log(s \vee n)/n}\} \leq 0.5$. \square

Lemma A.5. *Let $\mathcal{P}_{\nu-1}$ denote the set of all polynomial functions on $[0, 1]$ of degree $\nu - 1$. Assume conditions (C1), (C3) and (C5). Then, with probability approaching one, $\|h_j\|_{2,n}^2 \leq 1.5 \|h_j\|_2^2$ for all $h_j \in \mathcal{P}_{\nu-1}$ and $1 \leq j \leq d$.*

Proof. As in the previous proof, $C_{q,\nu}$ denotes some positive constant depending only on c_q and ν . Its value may change from line to line. By normalization, it suffices to show that $\max_{1 \leq j \leq d} \sup_{h_j \in \mathcal{H}_j} \|\|h_j\|_{2,n}^2 - 1\| \xrightarrow{P} 0$, where $\mathcal{H}_j := \{h_j : h_j \in \mathcal{P}_{\nu-1}, \|h_j\|_2 = 1\}$.

Pick any $1 \leq j \leq d$. By condition (C3) (ii) and Lemma A.1, $\|h_j\|_\infty \leq C_{q,\nu}\|h_j\|_2 = C_{q,\nu}$ for all $h_j \in \mathcal{H}_j$, and $\mathbb{E}[h_j(z_{1j})^4] \leq \|h_j\|_\infty^2 \|h_j\|_2^2 \leq C_{q,\nu}^2$. Therefore, by Massart's form of Talagrand's inequality, for all $t > 0$, with probability at least $1 - e^{-t}$,

$$Z_j \leq 2\mathbb{E}[Z_j] + C_{q,\nu}\sqrt{t/n} + C_{q,\nu}t/n,$$

where $Z_j := \sup_{h_j \in \mathcal{H}_j} |\|h_j\|_{2,n}^2 - 1|$. We wish to evaluate $\mathbb{E}[Z_j]$. Let $\sigma_1, \dots, \sigma_n$ be independent Rademacher random variables independent of $\mathbf{z}_1, \dots, \mathbf{z}_n$. By the symmetrization inequality and the contraction principle,

$$\mathbb{E}[Z_j] \leq C_{q,\nu}\mathbb{E}\left[\sup_{h_j \in \mathcal{H}_j} \left|\frac{1}{n} \sum_{i=1}^n \sigma_i h_j(z_{ij})\right|\right].$$

Arguing as in Meier et al. (2009, p.3813) (or using a standard entropy integral argument), it is shown that

$$\mathbb{E}\left[\sup_{h_j \in \mathcal{H}_j} \left|\frac{1}{n} \sum_{i=1}^n \sigma_i h_j(z_{ij})\right|\right] \leq \frac{C_{q,\nu}}{\sqrt{n}}.$$

Taking $t = 2 \log d$, we have, with probability at least $1 - d^{-2}$,

$$Z_j \leq C_{q,\nu}\sqrt{\frac{\log d}{n}}.$$

(Recall that $\log d/n \rightarrow 0$.)

By the union bound, the above inequality simultaneously holds for all $1 \leq j \leq d$ with probability at least $1 - d^{-1}$. Recalling that $d \rightarrow \infty$, we obtain the desired conclusion. \square

A.2 Proof of Lemma 6.1

Parts (i) and (ii): We first point out that (ii) follows from (i). Suppose that (i) is true for the case that u_i are independent Rademacher random variables independent of $\mathbf{z}_1, \dots, \mathbf{z}_n$. Let $\sigma_1, \dots, \sigma_i$ denote independent Rademacher random variables independent of $\mathbf{z}_1, \dots, \mathbf{z}_n$. By the symmetrization inequality for probabilities (van der Vaart and Wellner, 1996, Lemma 2.13), for all $t \geq \sqrt{8/n}$,

$$\mathbb{P}\left\{\sup_{g_j \in \mathcal{G}} \frac{|n^{-1} \sum_{i=1}^n \{g_j(z_{ij}) - \mu_{g_j}\}|}{\sqrt{\|g_j\|_2^2 + \epsilon^2 I(g_j)^2}} > t\right\} \leq 4\mathbb{P}\left\{\sup_{g_j \in \mathcal{G}} \frac{|n^{-1} \sum_{i=1}^n \sigma_i g_j(z_{ij})|}{\sqrt{\|g_j\|_2^2 + \epsilon^2 I(g_j)^2}} > \frac{t}{4}\right\}.$$

Thus, by (i), the right side goes to zero with $t = 4 \max\{\epsilon, C_1 \sqrt{\log(s \vee n)/n}\}$.

In what follows, we wish to show (i). Take C_ν as in Lemma A.3 and let $\epsilon = \epsilon_n \rightarrow 0$ be any sequence such that $\epsilon \geq C_\nu n^{-\nu/(2\nu+1)}$. Define the events

$$\begin{aligned}\Omega_7 &:= \{\|g_j\|_{2,n}^2 \leq 1.5\|g_j\|_2^2 + 0.5\epsilon^2 I(g_j)^2, \forall g_j \in \mathcal{G}, \forall j \in T\}, \\ \Omega_8 &:= \{\|h_j\|_{2,n}^2 \leq 1.5\|h_j\|_2^2, \forall h_j \in \mathcal{P}_{\nu-1}, \forall j \in T\},\end{aligned}$$

where $\mathcal{P}_{\nu-1}$ denotes the set of all polynomial functions on $[0, 1]$ of degree $\nu - 1$.

We first consider case (a) in condition (C2). By normalization, it suffices to consider the case that $|u_1| \leq 1$ almost surely. Pick any $j \in T$. Consider the function

$$\begin{aligned}F_j(\mathbf{u}, \mathbf{z}_j) &:= \sup_{g_j \in \mathcal{G}} \frac{|n^{-1} \sum_{i=1}^n u_i g_j(z_{ij})|}{\sqrt{\|g_j\|_2^2 + \epsilon^2 I(g_j)^2}}, \\ &= \sup_{g_j \in \mathcal{G}} \frac{n^{-1} \sum_{i=1}^n u_i g_j(z_{ij})}{\sqrt{\|g_j\|_2^2 + \epsilon^2 I(g_j)^2}}, \quad \mathbf{u} = (u_1, \dots, u_n)', \quad \mathbf{z}_j = (z_{1j}, \dots, z_{nj})',\end{aligned}$$

where the second inequality is due to the fact that \mathcal{G} is symmetric, i.e., if $g_j \in \mathcal{G}$ then $-g_j \in \mathcal{G}$. Given $\mathbf{z}_1, \dots, \mathbf{z}_n$, the map $\mathbf{u} \mapsto F_j(\mathbf{u}, \mathbf{z}_j)$ is Lipschitz continuous with Lipschitz constant bounded by

$$\sup_{g_j \in \mathcal{G}} \frac{n^{-1/2} \|g_j\|_{2,n}}{\sqrt{\|g_j\|_2^2 + \epsilon^2 I(g_j)^2}},$$

which is, on the event Ω_7 , bounded by $Cn^{-1/2}$. Therefore, by Corollary 4.8 of Ledoux (2001), on the event Ω_7 ,

$$P_u\{F_j(\mathbf{u}, \mathbf{z}_j) \geq E_u[F_j(\mathbf{u}, \mathbf{z}_j)] + Ctn^{-1/2}\} \leq c_1 e^{-c_2 t^2},$$

where P_u (E_u) denotes the conditional probability (expectation, respectively) with respect to u_1, \dots, u_n given $\mathbf{z}_1, \dots, \mathbf{z}_n$, and $c_1 > 0$ and $c_2 > 0$ are universal constants.

By Lemma A.2, invoke that $g_j \in \mathcal{G}$ can be written as $g_j = g_j^{[1]} + g_j^{[2]}$ such that (i) $\|g_j^{[1]}\|_\infty \leq \text{const.} \times I(g_j^{[1]})$ (the constant is independent of g_j); (ii) $g_j^{[2]} \in \mathcal{P}_{\nu-1}$; and (iii) $E[g_j^{[1]}(z_{1j})g_j^{[2]}(z_{1j})] = 0$. Observe that on the event $\Omega_7 \cap \Omega_8$,

$$\begin{aligned}\frac{|n^{-1} \sum_{i=1}^n u_i g_j(z_{ij})|}{\sqrt{\|g_j\|_2^2 + \epsilon^2 I(g_j)^2}} &\leq \frac{|n^{-1} \sum_{i=1}^n u_i g_j^{[1]}(z_{ij})|}{\sqrt{\|g_j^{[1]}\|_2^2 + \epsilon^2 I(g_j^{[1]})^2}} + \frac{|n^{-1} \sum_{i=1}^n u_i g_j^{[2]}(z_{ij})|}{\|g_j^{[2]}\|_2} \\ &\leq C \left\{ \frac{|n^{-1} \sum_{i=1}^n u_i g_j^{[1]}(z_{ij})|}{\sqrt{\|g_j^{[1]}\|_{2,n}^2 + \epsilon^2 I(g_j^{[1]})^2}} + \frac{|n^{-1} \sum_{i=1}^n u_i g_j^{[2]}(z_{ij})|}{\|g_j^{[2]}\|_{2,n}} \right\}.\end{aligned}$$

By Lemma A.3 (see also Addendum A.1) and the fact that u_i 's can be replaced by independent Rademacher variables by the contraction principle, the conditional expectation (given $\mathbf{z}_1, \dots, \mathbf{z}_n$) of the first term inside the brace is bounded by $C_\nu \epsilon$. On

the other hand, by Meier et al. (2009, p.3813), the conditional expectation of the second term inside the brace is bounded by a constant times $n^{-1/2}$ where the constant depends only on ν . So there exists a constant C'_ν depending only on ν such that $E_u[F_j(\mathbf{u}, \mathbf{z}_j)] \leq C'_\nu \epsilon$ on the event $\Omega_7 \cap \Omega_8$. Therefore, on the event $\Omega_7 \cap \Omega_8$,

$$P_u\{F_j(\mathbf{u}, \mathbf{z}_j) \geq C'_\nu \epsilon + C\sqrt{\log(s \vee n)/n}\} \leq c_1(s \vee n)^{-2},$$

where we have taken $t = \sqrt{2c_2^{-1} \log(s \vee n)}$.

We now move j . By the union bound,

$$\begin{aligned} & P\{\max_{j \in T} F_j(\mathbf{u}, \mathbf{z}_j) \geq C'_\nu \epsilon + C\sqrt{\log(s \vee n)/n}\} \\ & \leq P\left[\left\{\max_{j \in T} F_j(\mathbf{u}, \mathbf{z}_j) \geq \epsilon\right\} \cap \Omega_7 \cap \Omega_8\right] + P(\Omega_7^c) + P(\Omega_8^c) \\ & \leq c_1(s \vee n)^{-1} + P(\Omega_7^c) + P(\Omega_8^c). \end{aligned}$$

By Lemmas A.3 and A.4, $P(\Omega_7^c) + P(\Omega_8^c) \rightarrow 0$. Therefore, we obtain (i) for the case that $|u_1| \leq 1$ almost surely.

We next consider case (b) in condition (C2). Recall that $u_1|\mathbf{z}_1 \sim N(0, \sigma_u(\mathbf{z}_1)^2)$ and $\sigma_u(\mathbf{z}_1) \leq \sigma_u$ almost surely. Put $\tilde{u}_i := u_i/\sigma_u(\mathbf{z}_i)$. Then, $\tilde{u}_1, \dots, \tilde{u}_n$ are independent standard normal random variables independent of $\mathbf{z}_1, \dots, \mathbf{z}_n$. Consider now the function

$$F_j(\tilde{\mathbf{u}}, \mathbf{z}_1^n) := \sup_{g_j \in \mathcal{G}} \frac{|n^{-1} \sum_{i=1}^n \tilde{u}_i \sigma_u(\mathbf{z}_i) g_j(z_{ij})|}{\sqrt{\|g_j\|_2^2 + \epsilon^2 I(g_j)^2}}, \quad \tilde{\mathbf{u}} = (\tilde{u}_1, \dots, \tilde{u}_n)', \quad \mathbf{z}_1^n = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}.$$

Given $\mathbf{z}_1, \dots, \mathbf{z}_n$, the map $\tilde{\mathbf{u}} \mapsto F_j(\tilde{\mathbf{u}}, \mathbf{z}_1^n)$ is Lipschitz continuous with Lipschitz constant bounded by

$$\sigma_u \sup_{g_j \in \mathcal{G}} \frac{n^{-1/2} \|g_j\|_{2,n}}{\sqrt{\|g_j\|_2^2 + \epsilon^2 I(g_j)^2}},$$

which is, on the event Ω_7 , bounded by $C\sigma_u n^{-1/2}$. Therefore, by Theorem 7.1 of Ledoux (2001), on the event Ω_7 ,

$$P_{\tilde{\mathbf{u}}}\{F_j(\tilde{\mathbf{u}}, \mathbf{z}_1^n) \geq E_{\tilde{\mathbf{u}}}[F_j(\tilde{\mathbf{u}}, \mathbf{z}_1^n)] + C\sigma_u t n^{-1/2}\} \leq e^{-t^2/2}.$$

By the contraction principle for Gaussian processes (Ledoux and Talagrand, 1991, Corollary 3.17),

$$E_{\tilde{\mathbf{u}}}[F_j(\tilde{\mathbf{u}}, \mathbf{z}_1^n)] \leq C\sigma_u E_{\tilde{\mathbf{u}}} \left[\sup_{g_j \in \mathcal{G}} \frac{|n^{-1} \sum_{i=1}^n \tilde{u}_i g_j(z_{ij})|}{\sqrt{\|g_j\|_2^2 + \epsilon^2 I(g_j)^2}} \right].$$

The rest of the procedure is the same as the previous one.

Part (iii): The result basically follows from the same argument as the proof of Meier et al. (2009, Theorem 6). For the sake of completeness, we provide an outline of the proof. In what follows, $C_{q,\nu}$ denotes some constant depending only on c_q and ν . Its value may change from line to line.

Let $\sigma_1, \dots, \sigma_n$ denote independent Rademacher random variables independent of $\mathbf{z}_1, \dots, \mathbf{z}_n$. Define

$$Z := \sup_{g = \sum_{j=1}^d g_j, g_j \in \mathcal{G}} \frac{|||g||_{2,n}^2 - \|g\|_2^2|}{[\sum_{j=1}^d \sqrt{\|g_j\|_2^2 + \epsilon^2 I(g_j)^2}]^2},$$

$$\tilde{Z} := \sup_{g = \sum_{j=1}^d g_j, g_j \in \mathcal{G}} \frac{|n^{-1} \sum_{i=1}^n \sigma_i g(\mathbf{z}_i)|}{\sum_{j=1}^d \sqrt{\|g_j\|_2^2 + \epsilon^2 I(g_j)^2}}.$$

By Lemma A.1, for $g = \sum_{j=1}^d g_j, g_j \in \mathcal{G}$,

$$\|g\|_\infty \leq \sum_{j=1}^d \|g_j\|_\infty \leq C_{q,\nu} e^{-1/(2\nu)} \sum_{j=1}^d \sqrt{\|g_j\|_2^2 + \epsilon^2 I(g_j)^2},$$

and

$$\mathbb{E}[g(\mathbf{z}_1)^4] \leq \|g\|_\infty^2 \|g\|_2^2 \leq C_{q,\nu} e^{-1/\nu} \left[\sum_{j=1}^d \sqrt{\|g_j\|_2^2 + \epsilon^2 I(g_j)^2} \right]^4,$$

where we have use the inequality that $\|g\|_2 \leq \sum_{j=1}^d \|g_j\|_2 \leq \sum_{j=1}^d \sqrt{\|g_j\|_2^2 + \epsilon^2 I(g_j)^2}$. Thus, by Massart's (2000) form of Talagrand's (1996) inequality, for all $t > 0$, with probability at least $1 - e^{-t}$,

$$Z \leq 2\mathbb{E}[Z] + C_{q,\nu} \sqrt{\epsilon^{-1/\nu} t/n} + C_{q,\nu} \epsilon^{-1/\nu} t/n.$$

We wish to evaluate $\mathbb{E}[Z]$. Using the symmetrization inequality and the contraction principle, we have

$$\mathbb{E}[Z] \leq C_{q,\nu} \epsilon^{-1/(2\nu)} \mathbb{E}[\tilde{Z}].$$

By a standard calculation,

$$\mathbb{E}[\tilde{Z}] \leq \mathbb{E} \left[\max_{1 \leq j \leq d} \sup_{g_j \in \mathcal{G}} \frac{|n^{-1} \sum_{i=1}^n \sigma_i g_j(\mathbf{z}_{ij})|}{\sqrt{\|g_j\|_2^2 + \epsilon^2 I(g_j)^2}} \right].$$

Recall that $\delta = \max\{n^{-\nu/(2\nu+1)}, \sqrt{\log d/n}\}$. By lemma 13 of Meier et al. (2009) and

Lemma A.3, we have

$$\begin{aligned}
& \mathbb{E} \left[\max_{1 \leq j \leq d} \sup_{g_j \in \mathcal{G}} \frac{|n^{-1} \sum_{i=1}^n \sigma_i g_j(z_{ij})|}{\sqrt{\|g_j\|_2^2 + \epsilon^2 I(g_j)^2}} \right] \\
& \leq 4 \max_{1 \leq j \leq d} \mathbb{E} \left[\sup_{g_j \in \mathcal{G}} \frac{|n^{-1} \sum_{i=1}^n \sigma_i g_j(z_{ij})|}{\sqrt{\|g_j\|_2^2 + \epsilon^2 I(g_j)^2}} \right] + \frac{2C_{q,\nu} \epsilon^{-1/(2\nu)} (1 + \log d)}{3n} + \sqrt{\frac{4(1 + \log d)}{n}} \\
& \leq C_{q,\nu} \max \left\{ \epsilon, \frac{\epsilon^{-1/(2\nu)} \log d}{n}, \sqrt{\frac{\log d}{n}} \right\} \\
& \leq C_{q,\nu} \max\{\epsilon, \delta\}.
\end{aligned}$$

The last inequality is because

$$\frac{\epsilon^{-1/(2\nu)} \log d}{n} \lesssim \sqrt{\frac{\log d}{n^{2\nu/(2\nu+1)}}} \times \sqrt{\frac{\log d}{n}} = o(1) \sqrt{\frac{\log d}{n}}.$$

Thus, with probability at least $1 - e^{-t}$,

$$Z \leq C_{q,\nu} \max\{\epsilon^{(2\nu-1)/(2\nu)}, \epsilon^{-1/(2\nu)} \delta, \sqrt{\epsilon^{-1/\nu} t/n}, \epsilon^{-1/\nu} t/n\}.$$

Letting $t \rightarrow \infty$ sufficiently slowly (such that $\sqrt{t/n} \lesssim \epsilon$), we have, with probability approaching one, $Z \leq C_{q,\nu} \epsilon^{-1/(2\nu)} \max\{\epsilon, \delta\}$, which implies the desired conclusion. \square

B Proofs for Section 4

B.1 Proofs of Propositions 4.1 and 4.2

We first point out that since $\sum_{i=1}^n \tilde{\mathbf{x}}_i = \mathbf{0}$, by a standard argument, we may assume that $c^* = \mathbb{E}[y_1] = 0$ for the analysis of \hat{g} .

Proof of Proposition 4.1. The proof is a direct adaptation of that of Bickel et al. (2009, Theorem 6.1), so we omit the detail here. \square

For a subset $T \subset \{1, \dots, d\}$, let \mathbf{x}_{iG_T} denote the $m|T| \times 1$ vector stacked by \mathbf{x}_{iG_j} , $j \in T$.

Proof of Proposition 4.2. Recall that $\hat{T}^0 := \{j \in \{1, \dots, d\} : \|\hat{g}_j\|_{2,n} > 0\}$. Note that on the event Ω_0 , $\hat{T}^0 = \{j \in \{1, \dots, d\} : \|\hat{\beta}_{G_j}\|_E \neq 0\}$. By the Karush-Kuhn-Tucker condition, on the event Ω_0 ,

$$\frac{\sqrt{m}\lambda_1}{n} \frac{\hat{\Sigma}_j^{1/2} \hat{\beta}_{G_j}}{\|\hat{\Sigma}_j^{1/2} \hat{\beta}_{G_j}\|_E} = \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{x}}_{iG_j} (y_i - \tilde{\mathbf{x}}_i' \hat{\beta}), \quad \forall j \in \hat{T}^0,$$

which implies that on the event $\{\lambda_1 \geq 2\Lambda\} \cap \Omega_0$,

$$\begin{aligned}
\frac{\sqrt{m}\lambda_1}{n}\sqrt{\hat{s}} &\leq \left\| \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{x}}_{iG_{\hat{T}0}}(y_i - \tilde{\mathbf{x}}'_i \hat{\boldsymbol{\beta}}) \right\|_E \\
&\leq \left\| \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{x}}_{iG_{\hat{T}0}}(u_i + r_i) \right\|_E \quad (r_i := g^*(z_i) - \hat{g}(z_i)) \\
&\leq \left\| \frac{1}{n} \sum_{i=1}^n u_i \tilde{\mathbf{x}}_{iG_{\hat{T}0}} \right\|_E + \left\| \frac{1}{n} \sum_{i=1}^n r_i \tilde{\mathbf{x}}_{iG_{\hat{T}0}} \right\|_E \\
&\leq \frac{\sqrt{m}\lambda_1}{2n}\sqrt{\hat{s}} + 1.5 \left\| \frac{1}{n} \sum_{i=1}^n r_i \tilde{\mathbf{x}}_{iG_{\hat{T}0}} \right\|_E.
\end{aligned}$$

Applying the Cauchy-Schwarz inequality to the last line, we obtain

$$\frac{\sqrt{m}\lambda_1}{n}\sqrt{\hat{s}} \leq 3 \left\| \frac{1}{n} \sum_{i=1}^n r_i \tilde{\mathbf{x}}_{iG_{\hat{T}0}} \right\|_E \leq 3\|g^* - \hat{g}\|_{2,n} \hat{\phi}_{\max}(\hat{s}).$$

Thus, on the event $\{\lambda_1 \geq 2\Lambda \vee 0\} \cap \Omega_0$,

$$\hat{s} \leq \hat{C}^2 \hat{\phi}_{\max}(\hat{s})^2 s^*.$$

We wish to show that $\hat{\phi}_{\max}(\hat{s})^2 \leq \min_{s \in \mathcal{S}} \hat{\phi}_{\max}(s)^2$. Because the map $s \mapsto \hat{\phi}_{\max}(s)^2$ is non-decreasing, it suffices to show that $\hat{s} \leq s$ for any $s \in \mathcal{S}$. Pick any $s \in \mathcal{S}$. Suppose on the contrary that $\hat{s} > s$. Then,

$$\hat{s} \leq \hat{C}^2 \hat{\phi}_{\max}(\hat{s})^2 s^* = \hat{C}^2 \hat{\phi}_{\max}(s \cdot (\hat{s}/s))^2 s^* \leq \hat{C}^2 \lceil \hat{s}/s \rceil \hat{\phi}_{\max}(s)^2 s^* \leq 2\hat{C}^2 (\hat{s}/s) \hat{\phi}_{\max}(s)^2 s^* < \hat{s},$$

a contradiction (we have used the property $\hat{\phi}_{\max}(ls)^2 \leq \lceil l \rceil \hat{\phi}_{\max}(s)^2$ for $l > 0$, which can be shown in a similar way as the proof of Belloni and Chernozhukov (2011b, Lemma 8)). Therefore, we obtain the desired conclusion. \square

B.2 Proof of Theorem 4.2

Notation: In condition (C5), let K denote a fixed constant such that

$$\sup_{z \in [0,1]} \|(\psi_1(z), \dots, \psi_m(z))'\|_E \leq Km^{1/2}.$$

Theorem 4.2 follows from Propositions 4.1 and 4.2 together with Lemmas B.1-B.5 below. In what follows, we always assume conditions (C1)-(C5) and (C7).

Lemma B.1. *Assume that $m \log d/n \rightarrow 0$. Then, $P(\Omega_0) \rightarrow 1$.*

Proof of Lemma B.1. Without loss of generality, we may assume that $E[\mathbf{x}_1] = \mathbf{0}$. It suffices to show that $\max_{1 \leq j \leq d} \|\hat{\Sigma}_j - \mathbf{I}_m\| \xrightarrow{p} 0$. Because $\hat{\Sigma}_j = \hat{\Sigma}_{0j} - \bar{\mathbf{x}}_{G_j} \bar{\mathbf{x}}'_{G_j}$ ($\hat{\Sigma}_{0j} := n^{-1} \sum_{i=1}^n \mathbf{x}_{iG_j} \mathbf{x}'_{iG_j}$), it suffices to show that

$$\max_{1 \leq j \leq d} \|\bar{\mathbf{x}}_{G_j}\|_E \xrightarrow{p} 0, \max_{1 \leq j \leq d} \|\hat{\Sigma}_{0j} - \mathbf{I}_m\| \xrightarrow{p} 0.$$

The second assertion follows from Lemma 3.2 of Kato (2011). We wish to show the first assertion. Pick any $j \in \{1, \dots, d\}$. By Corollary 4.5 of Ledoux (2001), for all $t > 0$, with probability at least $1 - e^{-t^2/2}$, we have

$$\|\bar{\mathbf{x}}_{G_j}\|_E \leq (1 + tK) \sqrt{m/n},$$

where we have used the fact that $E[\|\bar{\mathbf{x}}_{G_j}\|_E] \leq \sqrt{n^{-2} \sum_{i=1}^n E[\|\mathbf{x}_{iG_j}\|_E^2]} = \sqrt{m/n}$ and $|\boldsymbol{\alpha}' \mathbf{x}_{1G_j}| \leq K\sqrt{m}$ (since we are considering the one-sided deviation inequality, 2 in front of the exponential term in Corollary 4.5 of Ledoux (2001) can be replaced by 1). By the union bound, the above inequality simultaneously holds for all $1 \leq j \leq d$ with probability at least $1 - de^{-t^2/2}$. Taking $t = 2\sqrt{\log d}$, we obtain the desired result. \square

Lemma B.2. *There exists a positive constant $A_{1,u}$ depending only on the distribution of u_1 such that for any λ_1 satisfying*

$$\lambda_1 \geq A_{1,u} \sqrt{n} \left(1 + \sqrt{\frac{\log d}{m}} \right),$$

we have $P\{\lambda_1 \geq 2\Lambda\} \rightarrow 1$.

Proof of Lemma B.2. This follows from deviation inequalities in product spaces. We first consider case (a) in condition (C2). By normalization, it suffices to consider the case that $|u_1| \leq 1$ almost surely. Pick any $1 \leq j \leq d$. Consider the function

$$F_j(\mathbf{u}, \mathbf{z}_{\cdot j}) := \left\| \sum_{i=1}^n u_i \tilde{\mathbf{x}}_{iG_j} / \sqrt{m} \right\|_E, \quad \mathbf{u} = (u_1, \dots, u_n)', \quad \mathbf{z}_{\cdot j} = (z_{1j}, \dots, z_{nj})'.$$

(Recall that \mathbf{x}_{iG_j} is generated by z_{ij} .) Given $\mathbf{z}_1, \dots, \mathbf{z}_n$, the map $\mathbf{u} \mapsto F_j(\mathbf{u}, \mathbf{z}_{\cdot j})$ is Lipschitz continuous with Lipschitz constant bounded by $\sqrt{n/m}$ (invoke that the maximum eigenvalue of $n^{-1} \sum_{i=1}^n \tilde{\mathbf{x}}_{iG_j} \tilde{\mathbf{x}}'_{iG_j}$ is 1). Thus, by Corollary 4.8 of Ledoux (2001), for all $t > 0$,

$$P_u\{F_j(\mathbf{u}, \mathbf{z}_{\cdot j}) \geq E_u[F_j(\mathbf{u}, \mathbf{z}_{\cdot j})] + Ct\sqrt{n/m}\} \leq c_1 e^{-c_2 t^2},$$

where P_u (E_u) denotes the conditional probability (expectation, respectively) of u_1, \dots, u_n given $\mathbf{z}_1, \dots, \mathbf{z}_n$, and $c_1 > 0$ and $c_2 > 0$ are universal constants. A direct calculation

shows that $\mathbb{E}_u[F_j(\mathbf{u}, \mathbf{z}_j)] \leq \sqrt{n}$ (invoke that the trace of the matrix $n^{-1} \sum_{i=1}^n \tilde{\mathbf{x}}_{iG_j} \tilde{\mathbf{x}}'_{iG_j}$ is bounded by m). Therefore, taking $t = \sqrt{2c_2^{-1} \log d}$, we have, with probability at least $1 - c_1 d^{-2}$,

$$F_j(\mathbf{u}, \mathbf{z}_j) \leq \sqrt{n} + C\sqrt{n \log d/m}.$$

By the union bound, the above inequality simultaneously holds for all $1 \leq j \leq d$ with probability at least $1 - c_1 d^{-1}$. Recalling that $d \rightarrow \infty$, we obtain the desired conclusion for the case that $|u_1| \leq 1$ almost surely.

In case (b) in Condition (C2), we use Theorem 7.1 of Ledoux (2001) instead of its Corollary 4.8. Recall that $u_1 | \mathbf{z}_1 \sim N(0, \sigma_u(\mathbf{z}_1)^2)$ and $\sigma_u(\mathbf{z}_1) \leq \sigma_u$ almost surely. Put $\tilde{u}_i := u_i / \sigma_u(\mathbf{z}_i)$. Then, $\tilde{u}_1, \dots, \tilde{u}_n$ are independent standard normal random variables independent of $\mathbf{z}_1, \dots, \mathbf{z}_n$. Define

$$F_j(\tilde{\mathbf{u}}, \mathbf{z}_1^n) := \left\| \sum_{i=1}^n \tilde{u}_i \sigma_u(\mathbf{z}_i) \tilde{\mathbf{x}}_{iG_j} / \sqrt{m} \right\|_E, \quad \tilde{\mathbf{u}} = (\tilde{u}_1, \dots, \tilde{u}_n)', \quad \mathbf{z}_1^n = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}'.$$

Given $\mathbf{z}_1, \dots, \mathbf{z}_n$, the map $\tilde{\mathbf{u}} \mapsto F_j(\tilde{\mathbf{u}}, \mathbf{z}_1^n)$ is Lipschitz continuous with Lipschitz constant bounded by $\sigma_u \sqrt{n/m}$. Thus, by Theorem 7.1 of Ledoux (2001), for all $t > 0$,

$$\mathbb{P}_{\tilde{\mathbf{u}}} \{F_j(\tilde{\mathbf{u}}, \mathbf{z}_1^n) \geq \mathbb{E}_{\tilde{\mathbf{u}}} [F_j(\tilde{\mathbf{u}}, \mathbf{z}_1^n)] + \sigma_u t \sqrt{n/m}\} \leq e^{-t^2/2},$$

where, by a direct calculation, $\mathbb{E}_{\tilde{\mathbf{u}}} [F_j(\tilde{\mathbf{u}}, \mathbf{z}_1^n)] \leq \sigma_u \sqrt{n}$. The rest of the procedure is exactly the same as the previous one. \square

Lemma B.3. *Let $\phi_{\max}(s)$ denote the s -th group sparse maximum eigenvalue of $\Sigma^{1/2}$: $\phi_{\max}(s) := \sup_{|T| \leq s, \alpha \in \mathbb{S}_T^{dm-1}} \|\Sigma^{1/2} \alpha\|_E$. Assume that s, m, d and n obey the growth condition*

$$\frac{s^2 m \log(d \vee n)}{n} \rightarrow 0. \quad (\text{B.1})$$

Then, $\hat{\phi}_{\max}(s) \lesssim_p 1$ provided that $\phi_{\max}(s) \lesssim 1$.

Recall that for a subset $T \subset \{1, \dots, d\}$, \mathbf{x}_{iG_T} denotes the $m|T| \times 1$ vector stacked by $\mathbf{x}_{iG_j}, j \in T$. We use the notation: $\hat{\Sigma}_T := n^{-1} \sum_{i=1}^n \tilde{\mathbf{x}}_{iG_T} \tilde{\mathbf{x}}'_{iG_T}$, $\hat{\Sigma}_{0T} := n^{-1} \sum_{i=1}^n \mathbf{x}_{iG_T} \mathbf{x}'_{iG_T}$ and $\Sigma_T := \mathbb{E}[\mathbf{x}_{1G_T} \mathbf{x}'_{1G_T}]$.

Proof of Lemma B.3. Without loss of generality, we may assume that $\mathbb{E}[\mathbf{x}_1] = \mathbf{0}$. Pick any subset T of $\{1, \dots, d\}$ such that $|T| = s$. We wish to evaluate a tail probability of $\|\hat{\Sigma}_T - \Sigma_T\|$. Since $\|\hat{\Sigma}_T - \Sigma_T\| \leq \|\hat{\Sigma}_{0T} - \Sigma_T\| + \|\tilde{\mathbf{x}}_{G_T}\|_E^2$, we separately evaluate the right two terms.

We first evaluate the term $\|\hat{\Sigma}_{0T} - \Sigma_T\|$. Invoke the expression

$$\|\hat{\Sigma}_{0T} - \Sigma_T\| = \sup_{\alpha \in \mathbb{S}^{sm-1}} \left| \frac{1}{n} \sum_{i=1}^n (\alpha' \mathbf{x}_{iG_T})^2 - \mathbb{E}[(\alpha' \mathbf{x}_{1G_T})^2] \right|.$$

Applying Massart's (2000) form of Talagrand's (1996) inequality to the right sum, for all $t > 0$, with probability at least $1 - e^{-t}$, we have

$$\|\hat{\Sigma}_{0T} - \Sigma_T\| \leq 2\mathbb{E}[\|\hat{\Sigma}_{0T} - \Sigma_T\|] + CK\phi_{\max}(s)\sqrt{tsm/n} + CK^2tsm/n,$$

where we have use the fact that $|\alpha' \mathbf{x}_{1G_T}| \leq K\sqrt{sm}$ and $\mathbb{E}[(\alpha' \mathbf{x}_{1G_T})^4] \leq K^2sm\mathbb{E}[(\alpha' \mathbf{x}_{1G_T})^2] \leq K^2sm\phi_{\max}(s)^2$. We now bound the expectation $\mathbb{E}[\|\hat{\Sigma}_{0T} - \Sigma_T\|]$ by Rudelson's (1999) inequality:

$$\mathbb{E}[\|\hat{\Sigma}_{0T} - \Sigma_T\|] \leq \max \left\{ CK\phi_{\max}(s)\sqrt{\frac{sm \log(sm)}{n}}, C^2K^2\frac{sm \log(sm)}{n} \right\}.$$

Thus, for all $t > 0$, with probability at least $1 - e^{-t}$, we have

$$\|\hat{\Sigma}_{0T} - \Sigma_T\| \leq CK \max \left\{ \phi_{\max}(s)\sqrt{\frac{sm(t \vee \log(sm))}{n}}, \frac{Ksm(t \vee \log(sm))}{n} \right\}.$$

Because the number of all subsets T of $\{1, \dots, d\}$ such that $|T| = s$ is $\binom{d}{s} \leq (ed/s)^s$, the above inequality simultaneously holds for all such T with probability at least $1 - \exp\{s \log(ed/s) - t\}$. Taking $t = 2s \log(ed/s)$ and recalling condition (B.1), we have

$$\max_{|T| \leq s} \|\hat{\Sigma}_{0T} - \Sigma_T\| = \max_{|T|=s} \|\hat{\Sigma}_{0T} - \Sigma_T\| \lesssim_p o(1)(\phi_{\max}(s) \vee 1) = o(1), \quad (\text{B.2})$$

provided that $\phi_{\max}(s) \lesssim 1$.

It remains to bound $\|\bar{\mathbf{x}}_{G_T}\|_E$. By Corollary 4.5 of Ledoux (2001), for all $t > 0$, with probability at least $1 - e^{-t^2/2}$, we have

$$\|\bar{\mathbf{x}}_{G_T}\|_E \leq (1 + CKt)\sqrt{sm/n},$$

where we have used the fact that $\mathbb{E}[\|\bar{\mathbf{x}}_{G_T}\|_E] \leq \sqrt{n^{-2} \sum_{i=1}^n \mathbb{E}[\|\mathbf{x}_{iG_T}\|_E^2]} = \sqrt{sm/n}$ and $|\alpha' \mathbf{x}_{1G_T}| \leq K\sqrt{sm}$. Taking $t = 2\sqrt{s \log(ed/s)}$, we have

$$\max_{|T| \leq s} \|\bar{\mathbf{x}}_{G_T}\|_E = \max_{|T|=s} \|\bar{\mathbf{x}}_{G_T}\|_E = o_p(1). \quad (\text{B.3})$$

Combining (B.2) and (B.3), we have

$$\hat{\phi}_{\max}(s)^2 \leq \phi_{\max}(s)^2 + \max_{|T| \leq s} \|\hat{\Sigma}_T - \Sigma_T\| = \phi_{\max}(s)^2 + o_p(1) \lesssim_p 1.$$

□

Lemma B.4. *Let κ denote the \mathbb{C} -restricted eigenvalue of $\Sigma^{1/2}$: $\kappa := \min_{\alpha \in \mathbb{S}^{dm-1} \cap \mathbb{C}} \|\Sigma^{1/2}\alpha\|_E$. Assume that s^*, m, d and n obey the growth condition*

$$\frac{(s^*)^2 m \log(d \vee n)}{n} \rightarrow 0. \quad (\text{B.4})$$

Then, we have $\hat{\kappa} \gtrsim_p 1$ provided that $\phi_{\max}(2s^) \lesssim 1$ and $\kappa \gtrsim 1$.*

Proof of Lemma B.4. We partly use an idea of Zhou (2009), but the overall proof is quite different. Put $c_0 = 21$. For any $\alpha \in \mathbb{C}$, we decompose α into a set of subvectors $\alpha_{G_{T_0}}, \alpha_{G_{T_1}}, \dots, \alpha_{G_{T_L}}$ in such a way that T_0 corresponds to the s^* largest groups in α in the Euclidean norm, T_1 corresponds to the s^* largest groups in $\alpha_{G_{T_0^c}}$, and so on. Then, we have $T_0^c = \cup_{l=1}^L T_l$ where $|T_l| = s^*$ for $1 \leq l \leq L-1$ and $|T_L| \leq s^*$. Because for each $l \geq 1$,

$$\|\alpha_{G_{T_l}}\|_E \leq \sqrt{s^*} \max_{j \in T_l} \|\alpha_{G_j}\|_E \leq \frac{1}{\sqrt{s^*}} \sum_{j \in T_{l-1}} \|\alpha_{G_j}\|_E,$$

we have

$$\begin{aligned} \sum_{l=1}^L \|\alpha_{G_{T_l}}\|_E &\leq \frac{1}{\sqrt{s^*}} \sum_{l=0}^{L-1} \sum_{j \in T_l} \|\alpha_{G_j}\|_E \\ &\leq \frac{1}{\sqrt{s^*}} \sum_{j=1}^d \|\alpha_{G_j}\|_E \\ &\leq \frac{1}{\sqrt{s^*}} (1 + c_0) \sum_{j \in T^*} \|\alpha_{G_j}\|_E \\ &\leq (1 + c_0) \|\alpha_{G_{T^*}}\|_E \\ &\leq (1 + c_0) \|\alpha_{G_{T_0}}\|_E, \end{aligned}$$

where we have use the fact that $\sum_{j \in T^*} \|\alpha_{G_j}\|_E \leq c_0 \sum_{j \in (T^*)^c} \|\alpha_{G_j}\|_E$, and $\|\alpha_{G_{T^*}}\|_E \leq \|\alpha_{G_{T_0}}\|_E$ by construction. Therefore, we have

$$\sum_{l=0}^L \|\alpha_{G_{T_l}}\|_E \leq (2 + c_0) \|\alpha_{G_{T_0}}\|_E. \quad (\text{B.5})$$

In what follows, we identify $\alpha_{G_{T_l}}$ as the $dm \times 1$ vector $\tilde{\alpha}$ such that $\tilde{\alpha}_{G_{T_l}} = \alpha_{G_{T_l}}$ and all the other elements of $\tilde{\alpha}$ are zero. Under this identification, α can be written as $\alpha = \sum_{l=0}^L \alpha_{G_{T_l}}$. Invoke now that

$$\begin{aligned} |\alpha'(\hat{\Sigma} - \Sigma)\alpha| &\leq \sum_{l=0}^L \sum_{l'=0}^L |\alpha'_{G_{T_l}}(\hat{\Sigma} - \Sigma)\alpha_{G_{T_{l'}}}| \\ &= \sum_{l=0}^L \|\alpha_{G_{T_l}}\|_E \sum_{l'=0}^L \|\alpha_{G_{T_{l'}}}\|_E \cdot |\mathbf{h}'_l(\hat{\Sigma} - \Sigma)\mathbf{h}_{l'}| \quad (\mathbf{h}_l := \alpha_{G_{T_l}} / \|\alpha_{G_{T_l}}\|_E) \\ &\leq \max_{0 \leq l, l' \leq L} |\mathbf{h}'_l(\hat{\Sigma} - \Sigma)\mathbf{h}_{l'}| \cdot \left(\sum_{l=0}^L \|\alpha_{G_{T_l}}\|_E \right)^2 \\ &\leq (2 + c_0)^2 \|\alpha_{G_{T_0}}\|_E^2 \max_{0 \leq l, l' \leq L} |\mathbf{h}'_l(\hat{\Sigma} - \Sigma)\mathbf{h}_{l'}|, \end{aligned} \quad (\text{B.6})$$

where we have used (B.5). Take and fix any l, l' such that $0 \leq l, l' \leq L$, and let $T' = T_l \cup T_{l'}$. It is not hard to see that

$$|\mathbf{h}'_l(\hat{\Sigma} - \Sigma)\mathbf{h}_{l'}| \leq \|\hat{\Sigma}_{T'} - \Sigma_{T'}\|.$$

Since $|T'| \leq 2s^*$, we see that

$$\max_{0 \leq l, l' \leq L} |\mathbf{h}'_l(\hat{\Sigma} - \Sigma)\mathbf{h}_{l'}| \leq \max_{|T| \leq 2s^*} \|\hat{\Sigma}_T - \Sigma_T\|. \quad (\text{B.7})$$

Combining (B.6) and (B.7), we have

$$|\hat{\kappa}^2 - \kappa^2| \leq \max_{\alpha \in \mathbb{S}^{dm-1} \cap \mathbb{C}} |\alpha'(\hat{\Sigma} - \Sigma)\alpha| \leq (2 + c_0)^2 \max_{|T| \leq 2s^*} \|\hat{\Sigma}_T - \Sigma_T\|.$$

By the proof of Lemma B.2, if (B.4) and $\phi_{\max}(2s^*) \lesssim 1$, we have $\max_{|T| \leq 2s^*} \|\hat{\Sigma}_T - \Sigma_T\| = o_p(1)$. Therefore, we have $\hat{\kappa}^2 \geq \kappa^2 - o_p(1) \gtrsim_p 1$ provided that $\kappa \gtrsim 1$. \square

Lemma B.5. *If $\|g^*\|_2^2 \lesssim s^*$, then $\inf_{g \in \tilde{\mathcal{G}}_m^{T^*}} \|g^* - g\|_{2,n}^2 \lesssim_p s^* \max\{m^{-2\nu}, n^{-1}\}$.*

Proof of Lemma B.5. Let g^m denote an element of $\mathcal{G}_m^{T^*}$ such that $\|g^* - g^m\|_2 = \inf_{g \in \mathcal{G}_m^{T^*}} \|g^* - g\|_2$. g^m can be written as $g^m(\mathbf{z}) = \sum_{j \in T^*} g_j^m(z_j)$ ($\mathbf{z} = (z_1, \dots, z_d)'$) and $g_j^m(\cdot) = c_j^* + \sum_{k=1}^m \beta_{jk}^* \psi_k(\cdot)$ for some $c_j^* \in \mathbb{R}, \beta_{jk}^* \in \mathbb{R}$ ($1 \leq k \leq m$) for each $j \in T^*$. Because $\mathbb{E}[g^*(\mathbf{z}_1)] = 0$, we have $\mathbb{E}[g^m(\mathbf{z}_1)] = 0$, so that each c_j^* may be taken such that $g_j^m(\cdot) = \sum_{k=1}^m \beta_{jk}^* (\psi_k(\cdot) - \mathbb{E}[\psi_k(z_{1j})])$. Take $\tilde{g}^m(\mathbf{z}) := \sum_{j \in T^*} \sum_{k=1}^m \beta_{jk}^* (\psi_k(z_j) - \bar{\psi}_{jk})$. Invoke now that

$$\begin{aligned} \|g^* - \tilde{g}^m\|_{2,n}^2 &\leq 2\|g^* - g^m\|_{2,n}^2 + 2\|g^m - \tilde{g}^m\|_{2,n}^2 \\ &= 2\|g^* - g^m\|_{2,n}^2 + 2\{n^{-1} \sum_{i=1}^n g^m(\mathbf{z}_i)\}^2. \end{aligned}$$

By condition (C7)-(c) and Markov's inequality, we have

$$\|g^* - g^m\|_{2,n}^2 \lesssim_p s^* m^{-2\nu},$$

while by the fact that $\mathbb{E}[g^m(\mathbf{z}_1)] = 0$, we have

$$\{n^{-1} \sum_{i=1}^n g^m(\mathbf{z}_i)\}^2 \lesssim_p n^{-1} \|g^m\|_2^2 \lesssim n^{-1} (\|g^*\|_2^2 + \|g^* - g^m\|_2^2) \lesssim s^* n^{-1}.$$

Therefore, we have

$$\inf_{g \in \tilde{\mathcal{G}}_m^{T^*}} \|g^* - g\|_{2,n}^2 \leq \|g^* - g^m\|_{2,n}^2 \lesssim_p s^* \max\{m^{-2\nu}, n^{-1}\}.$$

\square

B.3 Proof of Corollary 4.1

Take $g^m = \sum_{j \in T^*} g_j^m$ as in condition (C7)-(c)'. Without loss of generality, we may assume that $E[g_j^m(z_{1j})] = 0$ for all $j \in T^*$, so that each g_j^m is written as $g_j^m(\cdot) = \sum_{k=1}^m \beta_{jk}^0(\psi_k(\cdot) - E[\psi_k(z_{1j})])$ for some $\beta_{jk}^0 \in \mathbb{R}$ ($1 \leq k \leq m$). Define $\beta^* \in \mathbb{R}^{dm}$ by $\beta_{jk}^* = \beta_{jk}^0$ for $j \in T^*$ and $1 \leq k \leq m$, and $\beta_{G_j}^* = \mathbf{0}$ for $j \in (T^*)^c$. Let $\tilde{g}^m := \sum_{j=1}^d \tilde{g}_j^m$, $\tilde{g}_j^m(\cdot) := \sum_{k=1}^m \beta_{jk}^*(\psi_k(\cdot) - \bar{\psi}_{jk})$. By the proof of Lemma B.5, we have $\|g^* - \tilde{g}^m\|_{2,n}^2 \lesssim_p s^* m^{-2\nu}$. Invoke that

$$\|\hat{\beta} - \beta^*\|_E^2 \leq \hat{\phi}_{\min}(T^* \cup \hat{T})^{-2} \|\hat{g} - \tilde{g}^m\|_{2,n}^2 \lesssim_p \|\hat{g} - g^*\|_{2,n}^2 + \|g^* - \tilde{g}^m\|_{2,n}^2 \lesssim_p \frac{s^* m \lambda_1^2}{n^2},$$

so that

$$\|\sum_{j \in T^* \setminus \hat{T}^0} g_j^m\|_2^2 \leq \phi_{\max}(s^*)^2 \|\beta_{G_{T^* \setminus \hat{T}^0}}\|_E^2 \lesssim_p \frac{s^* m \lambda_1^2}{n^2}.$$

Therefore, we have

$$\|\sum_{j \in T^* \setminus \hat{T}} g_j^*\|_2^2 \leq 2\|\sum_{j \in T^* \setminus \hat{T}^0} g_j^m\|_2^2 + 2\|\sum_{j \in T^* \setminus \hat{T}^0} (g_j^* - g_j^m)\|_2^2 \lesssim_p \frac{s^* m \lambda_1^2}{n^2}.$$

□

C On condition (C2)

Suppose that z_1, \dots, z_n are given and fixed. As argued in Section 2.1, the key property of the error distribution to our rate analysis of Theorems 4.1 and 4.2 is the normal concentration property (around its mean) of a random variable of the form $\sup_{t \in \mathcal{T}} \sum_{i=1}^n u_i t_i$ where \mathcal{T} is a bounded and countable subset of \mathbb{R}^n , which means that, letting $Z := \sup_{t \in \mathcal{T}} \sum_{i=1}^n u_i t_i$,

$$P(Z \geq E[Z] + \sigma r) \leq C \exp(-cr^2), \quad \forall r > 0, \quad (\text{C.1})$$

where $\sigma^2 := \sup_{t \in \mathcal{T}} \sum_{i=1}^n t_i^2$, and $c > 0$ and $C > 0$ are fixed constants (see the proofs of Lemmas 6.1 and B.2). Condition (C2) gives a primitive sufficient condition for this normal concentration property. See Ledoux (2001) for an excellent exposition of the concentration of measure phenomenon. On the other hand, Meier et al. (2009) assumed a uniform subgaussian condition

$$E[\exp(u_1^2/L) | z_1] \leq M, \quad a.s., \quad (\text{C.2})$$

for some fixed constants $L > 0$ and $M > 0$, which is weaker than our condition (C2), but their established rate $s^*(\log d/n)^{2\nu/(2\nu+1)}$ is suboptimal. Suzuki et al. (2011) later

showed that the Meier et al. (2009) estimator achieves the minimax rate $s^*\delta^2$ but assumed that the error term is uniformly bounded.⁴ It is thus of some interest how our rate analysis changes if our condition (C2) is replaced by weaker (C.2).

To the best of the author's knowledge, it is not known whether the uniform subgaussian condition alone ensures the normal concentration property (C.1). However, by Theorem C.1 ahead, under the uniform subgaussian condition, a slightly weaker inequality

$$P(Z \geq E[Z] + \sigma r) \leq C \exp(-cr^2/\log n), \quad \forall r \geq 4, \quad (\text{C.3})$$

holds. A careful inspection of the proofs leads to that if condition (C2) is replaced by (C.2), under some modifications to the conditions, the rate of convergence of the second step estimator will be

$$\max \left\{ s^* n^{-2\nu/(2\nu+1)}, |\hat{T} \setminus T^*| \tilde{\delta}^2, \left\| \sum_{j \in T^* \setminus \hat{T}} g_j^* \right\|_2^2 \right\},$$

in the canonical case, where

$$\tilde{\delta} := \max \left\{ n^{-\nu/(2\nu+1)}, \sqrt{\frac{(\log n)(\log d)}{n}} \right\},$$

and rate of convergence of the group Lasso estimator will be $s^*\tilde{\delta}^2$. So the second step estimator at least achieves the rate $s^*\tilde{\delta}^2$. The only difference is the appearance of the additional $\log n$ term, and as long as $\log d/(n \log n) \rightarrow 0$, $s^*\tilde{\delta}^2$ is faster than the rate $s^*(\log d/n)^{2\nu/(2\nu+1)}$. It is also expected that, under the uniform subgaussian condition (C.2), the Meier et al. (2009) estimator has the same rate of convergence as $s^*\delta^2$.

C.1 Proof of (C.3)

Recall the ψ_α -norm:

$$\|X\|_{\psi_\alpha} = \inf\{s > 0 : E[\exp(X^\alpha/s^\alpha)] \leq 2\}, \quad \alpha > 0.$$

Let \mathcal{T} be a bounded and countable subset of \mathbb{R}^n .

Theorem C.1. *Let $\epsilon_1, \dots, \epsilon_n$ be independent random variables such that $\max_{1 \leq i \leq n} \|\epsilon_i\|_{\psi_2} \leq C_\psi$ for some constant C_ψ . Put $Z := \sup_{t \in \mathcal{T}} \sum_{i=1}^n \epsilon_i t_i$. Then, for all $r \geq 4$, we have*

$$P\{Z \geq E[Z] + r\sigma C_\psi\} \leq C \exp(-cr^2/\log n),$$

where $\sigma := \sqrt{\sup_{t \in \mathcal{T}} \sum_{i=1}^n t_i^2}$, and $c > 0$ and $C > 0$ are universal constants.

⁴Koltchinskii and Yuan (2010) and Raskutti et al. (2010) dealt with a different estimator and established the rate $s^*\delta^2$ under different settings. Koltchinskii and Yuan (2010) assumed that the error term is uniformly bounded, and Raskutti et al. (2010) assumed that the error term is normal independent of explanatory variables.

Theorem C.1 is essentially proved in Mendelson and Tomczak-Jaegermann (2008) (but the exponential term is slightly worse in Mendelson and Tomczak-Jaegermann (2008): $\exp(-cr^2/\log^2 n)$). For the sake of completeness, we provide a proof of the theorem. The proof of this theorem uses some properties of the ψ_1 -norm.

Lemma C.1. (*Ledoux and Talagrand, 1991, Theorem 6.21*) *Let ξ_1, \dots, ξ_n be independent centered random variables. Then,*

$$\left\| \sum_{i=1}^n \xi_i \right\|_{\psi_1} \leq C \left(\mathbb{E} \left[\left\| \sum_{i=1}^n \xi_i \right\| \right] + \left\| \max_{1 \leq i \leq n} |\xi_i| \right\|_{\psi_1} \right),$$

where C is a universal constant.

For the evaluation of the term $\left\| \max_{1 \leq i \leq n} |\xi_i| \right\|_{\psi_1}$, we use the next lemma.

Lemma C.2. (*van der Vaart and Wellner, 1996, Lemma 2.2.2*) *Let ξ_1, \dots, ξ_n be any random variables. Then,*

$$\left\| \max_{1 \leq i \leq n} |\xi_i| \right\|_{\psi_1} \leq C(\log n) \max_{1 \leq i \leq n} \|\xi_i\|_{\psi_1},$$

where C is a universal constant.

Proof of Theorem C.1. In this proof, c and C denote some universal constants. Their values may change from line to line. Let $\epsilon_i^- := \epsilon_i I(|\epsilon_i| \leq L)$ and $\epsilon_i^+ := \epsilon_i I(|\epsilon_i| > L)$. The constant $L > 0$ is defined later. Define $\tilde{\mathcal{T}} := \mathcal{T} \cup \{-t : t \in \mathcal{T}\}$. Let $Z^- := \sup_{t \in \mathcal{T}} \sum_{i=1}^n \epsilon_i^- t_i$ and $\check{Z}^+ := \sup_{t \in \tilde{\mathcal{T}}} \sum_{i=1}^n \epsilon_i^+ t_i$. Clearly, $Z \leq Z^- + \check{Z}^+$, and by using that $\sum_{i=1}^n \epsilon_i^- t_i = \sum_{i=1}^n (\epsilon_i - \epsilon_i^+) t_i = \sum_{i=1}^n \epsilon_i t_i + \sum_{i=1}^n \epsilon_i^+ (-t_i)$, we have $\mathbb{E}[Z^-] \leq \mathbb{E}[Z] + \mathbb{E}[\check{Z}^+]$, so that $\mathbb{E}[Z] \geq \mathbb{E}[Z^-] - \mathbb{E}[\check{Z}^+]$. Observe that

$$\begin{aligned} \mathbb{P}\{Z \geq \mathbb{E}[Z] + r\sigma C_\psi\} &\leq \mathbb{P}\{Z^- + \check{Z}^+ \geq \mathbb{E}[Z^-] - \mathbb{E}[\check{Z}^+] + r\sigma C_\psi\} \\ &\leq \mathbb{P}\{Z^- \geq \mathbb{E}[Z^-] + r\sigma C_\psi/2\} + \mathbb{P}\{\check{Z}^+ + \mathbb{E}[\check{Z}^+] \geq r\sigma C_\psi/2\}, \end{aligned}$$

Because $|\epsilon_i^-| \leq 2L$, by Corollary 4.8 of Ledoux (2001), we have

$$\mathbb{P}\{Z^- \geq \mathbb{E}[Z^-] + r\sigma C_\psi/2\} \leq C \exp(-cr^2 C_\psi^2 / L^2), \quad \forall r > 0.$$

On the other hand, it is standard to see that $\|(\epsilon_i^+)^2\|_{\psi_1} = \|\epsilon_i^+\|_{\psi_2}^2 \leq C_\psi^2$ and

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^n (\epsilon_i^+)^2\right] &\leq n \max_{1 \leq i \leq n} \mathbb{E}[\epsilon_i^2 I(|\epsilon_i| > L)] \\ &\leq n \max_{1 \leq i \leq n} \mathbb{E}[\epsilon_i^4]^{1/2} \mathbb{P}(|\epsilon_i| > L)^{1/2} \\ &\leq n C C_\psi^2 \exp(-cL^2 / C_\psi^2). \end{aligned}$$

Thus, by Lemmas C.1 and C.2,

$$\begin{aligned} \left\| \sum_{i=1}^n (\epsilon_i^+)^2 \right\|_{\psi_1} &\leq \left\| \sum_{i=1}^n \{(\epsilon_i^+)^2 - \mathbb{E}[(\epsilon_i^+)^2]\} \right\|_{\psi_1} + \mathbb{E} \left[\sum_{i=1}^n (\epsilon_i^+)^2 \right] \\ &\leq CC_\psi^2 \{n \exp(-cL^2/C_\psi^2) + \log n\}. \end{aligned}$$

Take $L = CC_\psi \sqrt{\log n}$ such that $\mathbb{E}[\sum_{i=1}^n (\epsilon_i^+)^2] \leq C_\psi^2$ and $\left\| \sum_{i=1}^n (\epsilon_i^+)^2 \right\|_{\psi_1} \leq CC_\psi^2 \log n$. Because $\check{Z}^+ \leq \sigma \{\sum_{i=1}^n (\epsilon_i^+)^2\}^{1/2}$, we have

$$\|\check{Z}^+\|_{\psi_2} \leq C\sigma C_\psi \sqrt{\log n}, \quad \mathbb{E}[\check{Z}^+] \leq \sigma C_\psi.$$

Therefore, for all $r \geq 4$, we have

$$\begin{aligned} \mathbb{P}\{\check{Z}^+ + \mathbb{E}[\check{Z}^+] \geq r\sigma C_\psi/2\} &\leq \mathbb{P}\{\check{Z}^+ \geq r\sigma C_\psi/4\} \\ &\leq C \exp(-cr^2/\log n). \end{aligned}$$

This completes the proof. □